THE UNIVERSITY
of EDINBURGH

Development of an information theory based computational framework for the analysis of molecular dynamics simulations of proteins under allosteric regulation

Lisa Patrick

Doctor of Philosophy

University of Edinburgh

2019

Declaration of Authorship

I, Lisa PATRICK, declare that this thesis titled, "Development of an information theory based computational framework for the analysis of molecular dynamics simulations of proteins under allosteric regulation" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"It has been said that everything everywhere affects everything else. This may be true. Or perhaps the world is just full of patterns."

Sir Terry Pratchett

Abstract

Allosteric signalling was first discovered over 50 years ago, yet the underlying molecular determinants are not yet completely understood. The ability to predict the activity of allosteric small molecules could have a huge therapeutic impact, as targeting allosteric sites in proteins potentially presents significant benefits over active site inhibitors, in both selectivity and efficacy. While some systems undergo fairly well understood structural changes, there is no overall model that satisfactorily describes how allostery works. Molecular dynamics (MD) simulations provide a tool to study protein dynamics at the atomistic level, however traditionally employed analysis methods have proven inadequate to deliver a mechanistic description of allostery, which can be applied broadly to a range of allosteric systems.

This thesis presents the development of a Python workflow for the analysis of Molecular Dynamics (MD) simulations of proteins subjected to allosteric regulation. The end goal is to provide a new tool for structure-based drug design (SBDD) for these systems. This tool computes various descriptors, such as distances, torsions, collective motions and interaction energies, and then utilises two concepts from information theory to compare these descriptors: Kullback-Leibler (KL) divergence and Mutual Information (MI). MI is used to determine correlation between simulation descriptors that can aid explanation of conformation/activity relationships; while KL divergence is used to highlight differences of one descriptor between simulations of related molecular systems.

Proof of concept for this approach utilises the protein phosphoinositide dependent kinase-1 (PDK1) as a test case. This protein plays a crucial role in cell signalling, by activation of other kinases within the same

family (AGC kinases). Inhibition of PDK1 has been of much interest, as over-expression and dysfunction is related to several diseases, most notably cancer. Active site compounds suffer from selectivity issues, as the active site is well conserved across all AGC kinases, however PDK1 has a well defined allosteric site, with known peptide and small molecule activators and inhibitors. Therefore understanding this mechanism could facilitate design of more selective allosteric drugs. Long MD trajectories were run for PDK1 in complex with three different drug like molecules for which crystallographic data was available: two activators, and one inhibitor. In order to mimic experimental assay conditions, simulation systems were composed of PDK1, the covalently bound allosteric small molecule, ATP, two Mg^{2+} ions, a model of a substrate peptide, and a box of explicitly modelled water molecules. Simulations were performed with the software Sire/OpenMM Molecular Dynamics (SOMD). From the resulting trajectories, the KL analysis workflow was able to identify conformational differences between the activated and inhibited systems, and identify the dominant motions leading to these structural changes. Subsequently, an energetic comparison was performed using a per-residue decomposition of the non-bonded interactions between different components of the system (protein, ligand, ATP and substrate). Calculating MI of these energies relative to structural features highlighted that the motion of the activation loop in PDK1 is highly correlated with the interaction energy of ATP with the protein only when an allosteric ligand is bound. Further evidence to support this observation was obtained using an extended set of 21 further compounds for which activity data was available, which share the same scaffolds as the two activators initially studied. This confirms there is a unique conformation of the activation loop achieved only by the highest activating compounds, and not by the inhibited complex, and that this is correlated with the interactions of the protein with ATP.

To extend the applicability of this methodology, our attention shifted to the more challenging test case posed by protein-tyrosine phosphatase 1B (PTP1B). PTP1B is a promising target for the treatment of obesity and diabetes, as mice with deletion of the PTPN1 gene (which encodes PTP1B)

show significant resistance to both conditions. As with PDK1, the active site of protein tyrosine kinases is well conserved, and so selective phosphotyrosine analogue inhibitors, which bind at the active site, are difficult to develop. In this case, exploration of the key conformational changes required the use of enhanced sampling techniques, as these processes occur on millisecond timescales, and therefore cannot easily be sampled using equilibrium MD. In particular, steered-MD simulations were needed to probe the movements of the “WPD” loop, which closes over the substrate during the catalytic cycle, and positions key residues to interact with the substrate. The allosteric inhibitors for this system are believed to stabilise the "open" loop conformation, and restrict the loop closing into the active conformation. Therefore understanding how this stabilisation occurs is crucial in order to design more effective inhibitors. From the initial steered-MD run, a “swarm of trajectories” approach was applied, seeding hundreds of equilibrium MD runs from intermediate structures gathered during the steered-MD. This was used to generate a Markov State Model description of the conformational changes involved, in order to compare the loop closing mechanism for the inhibitor-bound, and substrate-bound simulations. This generates intermediate states of the loop closing, where KL can highlight structural differences between the states.

Overall, this work provides a generally applicable toolkit for the analysis of equilibrium and biased MD simulations to predict and characterise allosteric coupling in protein structural ensembles.

Lay summary

The work in this thesis describes the development of computational methods to analyse simulations of proteins, with a view to understanding how a particular biological process known as allostery works in order to utilise it for drug design.

In general the goal of drug design is to identify a particular biomolecule, usually a protein or enzyme, which is involved in a disease pathway, and design a molecule which can bind to the surface of the biomolecule, and in doing so change its behaviour. Such a molecule is referred to as a ligand. This is a desirable effect if a biomolecule is malfunctioning, or can have some knock on effect to another biological system, where altering the function will have some therapeutic effect. Usually this is achieved by designing a ligand which binds to the "active site", the region of the protein that is responsible for the function, usually the site of catalysis for an enzyme, and by binding a ligand there, the normal process is directly blocked. However it is also possible to bind to a region located away from the active site, and as this binding affects the conformation and dynamical behaviour of the protein as a whole, it is possible to alter the function at the active site, even though the ligand is bound to a different region of the protein. This distant effect is known as allostery, and it is an important mechanism in normal cell regulation and proliferation, as it is used as a signalling mechanism. Designing drugs that work in this way has many benefits in selectivity and efficacy over active site ligands, but the process is not well enough understood to make routine design of these molecules simple.

We can study this process using Molecular Dynamics (MD) simulations; a technique which uses Newton's classical equations of motion to simulate the motions of atoms and molecules. By simulating a protein without any allosteric ligand bound, and then again with different allosteric ligands, we

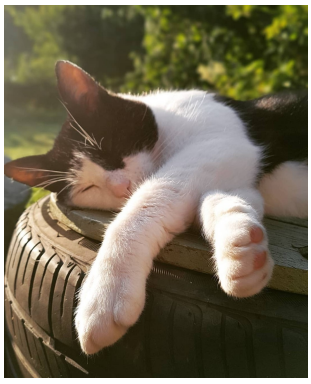
can compare the motions between the non-ligand-bound, and ligand-bound simulations, to highlight the effects that the allosteric ligand causes. To do this we use different mathematical and statistical techniques to analyse the data we obtain from the MD simulation. The first technique is known as PCA (Principal component analysis) which reduces the complexity of looking at the motions of thousands of atoms, by implementing a linear combination of these motions into a fewer number of "collective motions", where these new motions still represent the majority of the variability in the data. This allows us to highlight the largest collective motions (highest variance) of the protein, and in many cases these large scale motions are of biological significance.

We can also then look at other structural differences such as torsional angles, interaction energies, or distances within the protein. From our MD simulation we have many snapshots of the system, so we can make many of each of these measurements as the protein moves. From these measurements we can make distributions, and then compare the distributions from the different simulations using a concept called the Kullback-Leibler divergence, which comes from a branch of mathematics known as Information Theory. This tells us how different a particular measurement is between the "protein only" simulation, and the simulation with the protein plus the allosteric ligand.

Finally, we can then use another concept from Information Theory known as Mutual Information, to tell us how correlated two variables are. This allows us to determine whether regions in different areas of the protein show correlated motions, or whether motions correlate with interaction energies relating directly to the reaction mechanism, and this could help to explain allostery.

Acknowledgements

Firstly I would like to thank my supervisor, Dr Julien Michel, for your support, guidance and insights throughout all stages of this project. I would also like to thank Ben Cossins at UCB, for your input to the project and ensuring my placements at UCB were extremely enjoyable. From the Michel group, I would like to thank particularly Dr Jordi Juarez-Jimenez and Dr Antonia Mey, as without their support and good banter I think I might not have made it through these 4 years. You have answered endless questions on proteins and coding and managed to remain extraordinarily patient despite many stupid questions. Also I would like to thank all members of the Michel group, past and present, and all other occupants of the computational office for making this office a lovely place to work. Special mention goes to Georgia for the good friendship, and nice trips to drink whisky; to Rui for the extensive computer support; and to Darren, for his exhaustive supply of truly awful jokes. It wouldn't have been the same without them. Last but not least, I'd like to thank my partner Derek, who has tolerated my (many) moments of stress and long thesis writing days, you have been incredibly supportive. And final small mention to Edgar, my best (non-human) friend.



Contents

List of Figures

List of Tables

List of Abbreviations

AGC	Related to protein kinases **A**, **G** and **C**
Amber	Assisted Model Building with Energy Refinement
aPKs	Atypical Protein Kinases
CAMK	Ca^{2+} / calmodulin-dependent kinases
CK1	Casein kinase 1
CMGC	**C**yclin-dependent kinases, **M**itogen-activated protein kinases, **G**lycogen synthase kinases, **C**yclin-dependent kinase-like kinases
DSKs	Dual-Specificity Kinases
ePKs	Eukaryotic Protein Kinases
GAFF	General AMBER force field
Gromacs	GROningen MAchine for Chemical Simulations
HM	Hydrophobic motif
KL	Kullback-Leibler
MD	Molecular Dynamics
MFPTs	Mean First Passage Times
MI	Mutual Information
PDK1	Phosphoinositide-dependent kinase-1
PPI	Protein-protein Intereface
PSPs	Protein Ser/Thr Phosphatases
PTP1B	Tyrosine-protein phosphatase non-receptor type 1
PTPN1	Protein Tyrosine Phosphatase Non-Receptor Type 1
PTPs	Protein Tyr Phosphatases
RGC	Receptor guanylyl cyclase
SBDD	Structure Based Drug Design
sMD	Steered Molecular Dynamics
STKs	Serine/Threonine Kinases
TKs	Tyrosine kinases

xxx

Chapter 1

Introduction

1.1 Introduction

The term "allostery" was first presented in work by Jacques Monod in 1961 [1, 2], to describe proteins which have more than one ligand binding site, and where binding at one site results in a change in activity at the other.

> *"These proteins are assumed to possess two, or at least two, stereospecifically different, non-overlapping receptor sites. One of these, the actice site, binds the substrate and is responsible for the biological activity of the protein. The other, or allosteric site, is complementary to the structure of another metabolite, the allosteric effector, which it binds specifically and reversibly."* [2]

Figure 1.1 summarises the two initial models presented to describe allostery: the Monod-Wyman-Changeux (MWC) model [3] and the Koshland-Némethy-Filmer (KNF) model [4]. For a multimeric system, binding of a ligand to one subunit can affect the ligand binding affinity of another subunit. The MWC model describes this cooperativity with a two state model, where binding of a ligand to one subunit shifts the equilibrium of these two states, and this is the explanation for the cooperativity seen. The KNF model on the other hand considers ligand binding steps to be sequential: in that binding of a ligand to one subunit instigates a conformational change in that subunit, which affects neighbouring subunit conformations and affects their binding affinity.

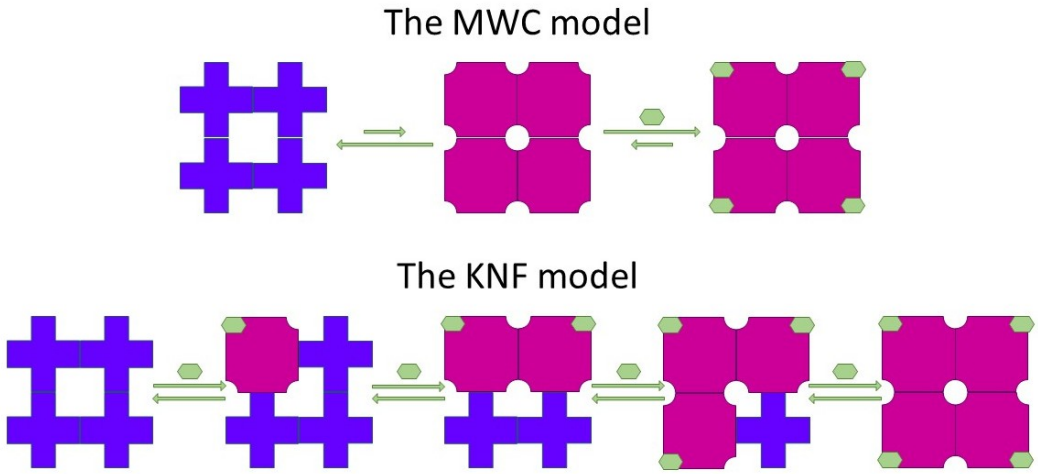


FIGURE 1.1: The MWC model of protein allostery which describes two states, termed as tense (purple) or relaxed (magenta), either of which can be stabilised by allosteric binding of a ligand (green).

These two, and other early models, based their descriptions on the study of hemoglobin, where cooperativity between the 4 subunits that form the quaternary complex is essential for function. This phenomenon was known for some time before the concept of allostery was developed, as studies by Christian Bohr in 1904 [5] found that after the first molecule of oxygen binds, the subsequent binding to the three remaining subunits occurs more readily. This cooperativity can be seen by noting that the binding affinity does not increase linearly with ligand concentration, and this can be either positive or negative; where initial binding promotes further binding, or restricts further binding. This effect is allostery: binding of one oxygen molecule the hemoglobin tetramer has an effect in another subunit at a site distant from the initial binding site. Hemoglobin also exhibits both "homotropic" and "heterotropic" allostery: heterotropic in that binding of oxygen promotes binding of the same molecule at another site; and homotropic as binding 2,3-bisphosphoglyceric acid allosterically decreases oxygen affinity at other subunits [6]. The effect of pH on oxygen affinity to hemoglobin is also an allosteric effect, as protonation facilitates salt bridge formation that stabilises the low oxygen affinity state. So the importance of allostery on important biological mechanisms has been known for some time, and so utilising this

for drug design is of considerable therapeutic interest.

Most early models also consider allostery to be a property of multimeric systems. But allostery is ubiquitous throughout the proteome; and in fact it has been suggested that *all* proteins could be allosteric [7]. This seems quite possible, since a vast array of processes utilise allostery for regulation and control of biological mechanisms, and is crucial for both long and short range signalling pathways [8]. The importance of correct cell signalling cannot be over-emphasised: cells must be able to respond to very subtle changes in environment, or face problems with cell development, cell growth, and cell survival. Ultimately all of these responses to environment are a result of allosteric effects. The definition of allostery also encompasses various forms of allosteric "trigger" as was discussed with hemoglobin: this could be homotropic or heterotropic allostery by binding of a small molecule; binding of an ion; binding of another protein; a mutation; or in response to light [9] or pH changes. Allosteric mechanisms are responsible for many regulatory processes and information transfer, over a wide range of important biological processes, yet a detailed description of the mechanism involved has proven difficult. On binding of a ligand in a region which is distinct from the active site, termed the allosteric site, some conformational or dynamic change then ensues, which affects the binding of another molecule at the active, or orthosteric site.

How this signal is transmitted from one site to another has been the topic of much research [10, 11], and yet sufficient information is not yet available to be able to design drugs to routinely target these sites. While some systems have fairly well understood structural changes [12], there has been no overall description [13], which can applied over a range of allosteric proteins and be used to predict the effects of ligand binding.

1.2 Structure based drug design

In the 1950s and 60s the first crystal structures of proteins were obtained by Max Perutz and John Kendrew, who developed the method of protein crystallography and solved the structure of myoglobin [14], and then hemoglobin

shortly after. Around a decade later, they deposited the first PDB file in the protein data bank, structure 1MBN. The insights they gained from analysis of these structures helped to develop understanding of sickle cell anaemia, and won them a Nobel prize in 1962. Since then, the availability of 3D structures of proteins has expanded massively, with now over 150k structures available in the protein data bank. This has been instrumental to the development of our understanding of protein function, and ability to design drugs. Understanding of structure and function of proteins is extremely important, however the crucial aspect for drug design is that crystal structures also give information about the location and environment of ligand binding sites, and this is the basis of structure based drug design (SBDD).

Computer aided drug design facilitates the ability to rationally design drugs, and has now become a routinely used method in drug discovery. Initial methods relied on use of static structures, for example docking [15], or using scoring functions [16]. Molecular simulation allows for the study of dynamics of proteins at atomistic level. Accuracy and speed of simulations of protein dynamics has seen rapid advances in the last decade, and it is now possible to simulate large systems to reasonably long timescales (μs to ms) [17].

1.3 Drug design and allostery

In the last century, the vast amount of efforts in drug design have been aimed at small molecules which bind at the active site of proteins. This could be to inhibit enzyme function, to alter the behaviour of receptors, or membrane channels [18, 19]. However often problems of selectivity are an issue, and some particular drug targets were deemed "undruggable" due to lack of suitable selective inhibitors.

Allostery presents an untapped resource of potential drug targets. Binding a drug at a site which is distant to the active site has numerous advantages in both selectivity and efficacy. It is often the case that the active

site across families of proteins are similar: a result of evolutionary pressure. Conservation of residues at the active site makes a lot of sense; enzymes carry out a very specific function, and are often highly specific in the substrates they bind. Any mutations or variation of the residues directly involved in catalysis or substrate binding would likely have a negative outcome on functionality, and are therefore not tolerated. This makes selective drugs extremely difficult to develop. Binding sites distant to the catalytic centre are far more likely to tolerate mutations without affecting overall function, so allosteric sites have far more potential to be selective. In addition, that the ability to tune inhibition, or even activate function of enzymes, will give allosteric drugs the edge over traditional active site compounds. Furthermore, this process can be either reversible or irreversible: since allosteric compounds can partially inhibit or activate function, it is even possible to include covalent drugs, which would bind irreversibly to their targets. Yet holding back development of these drugs is the limitations of our understanding of allosteric mechanisms, and progress will need to be made before routine, rational design, will be possible. Predictive tools which can guide compound design are essential.

Allostery has historically been understood as a structural phenomenon, in that binding *induces* some conformational change. But conformational selection is far more in line with our current understanding of protein behaviour. This is where the MWC model was in part correct, although describing only a two state system, the idea of conformational selection was apparent. The result of binding is a shift in equilibrium, rather than a structural change induced by binding.

This seems much more reasonable an explanation for allostery. As it is well understood that protein behaviour is statistical in nature, where many conformations of the protein are accessible, and the distribution of conformational states define a conformational ensemble around the native state [20]. Allostery is a result of shifting of states: where binding of a ligand at a site distant to the active site favours a particular conformational state,

and results in shifting of the population of these states by pushing the equilibrium in that direction. Allostery can be seen without significant conformational change, and many studies, even as early as the 1980's [21] have shown dynamical changes only as a method of allosteric regulation. If an allosteric modulator selects a particular conformation for binding, how can these subtle changes relate to changes in activity, and how can we understand this process well enough to exploit it for drug design? It is also very possible that allostery is a combination of both conformational selection and induced fit [22], where a particular conformation selected for binding is then stabilised, and allows for smaller adjustments to reach the catalytically active state after binding [23–25].

Molecular dynamics (MD) simulations are routinely used to investigate protein behaviour in atomistic detail. Understanding protein dynamics from vast quantities of simulation data is already extremely difficult: add then the complexity and diversity of allosteric mechanisms, and the challenge becomes even harder. For a MD simulation of a protein, motions of hundreds of residues (thousands of atoms), are computed. How can we distinguish relevant motions or functionally useful interactions relevant to allostery, from stochastic, non-functional thermal fluctuations? Development of tools which can extract useful information are important, as is the ability to relate these dominant motions and important structural features to actual differences in activity. To understand allostery, difficulty also lies in relating how particular local, and often subtle structural change, results in global functional changes, and how the specific structure of a ligand can facilitate this. It is not always clear what structural differences can be observed between the inactive and active conformations of a protein, and can be even more difficult when considering varying degrees of activity (i.e. strong activators vs. weak activators), so methods to highlight subtle differences must be available. Considering allostery under the framework of conformational selection; we must also understand how a particular conformation is best suited for allosteric drug binding.

Yet allosteric sites present an opportunity for selective drugs, in particular for highly sought after targets such as kinases, GPCRs (G protein-coupled receptors) or extremely challenging or otherwise "undruggable" targets such PTPs (Protein Tyrosine Phosphatases); therefore development of methods to improve rational design are crucial. Allostery also offers "tunable" modulators. While active site inhibitors can only inhibit, and can only completely switch off the function of the protein, allosteric modulators can activate (to varying degrees), inhibit, or even partially inhibit. The main issues with development of allosteric drugs is due to an incomplete mechanistic understanding of allostery, and how particular molecules achieve varying degrees of allosteric activation or inhibition. Furthermore, allosteric sites are often far more shallow than active sites, as many are usually protein-protein interfaces (PPIs) and therefore do not have the same potential to bind a small molecule with high affinity than the active site. Therefore it is extremely important to have methods which can not only identify effective allosteric sites on proteins, but also gain some mechanistic understanding to allow rational design of compounds that bind to these sites and have the desired effect on activity.

Many approaches have been developed to tackle these challenges; and is is likely the case that we require many different parallel approaches in order to fully understand allostery. In the last 15 or so years, there has been a surge in the interest in allosteric sites, which can be seen in figure 1.2, however the number is still small relative to the work done on active site compounds. Many methods to attempt to define an underlying mechanism for allostery, or to predict allosteric effects have been developed. Often, computational methods rely on analysis of MD trajectories, followed by some form of dimensionality reduction, such as Principal Component Analysis (PCA) [26–29]. PCA aims to explain the majority of variability in the data within in a few collective modes, rather than the 3N dimensions for N atoms. Alternatives to MD include simplification of the conformational energy surface using techniques such as Normal mode analysis (NMA) [30]. NMA relies

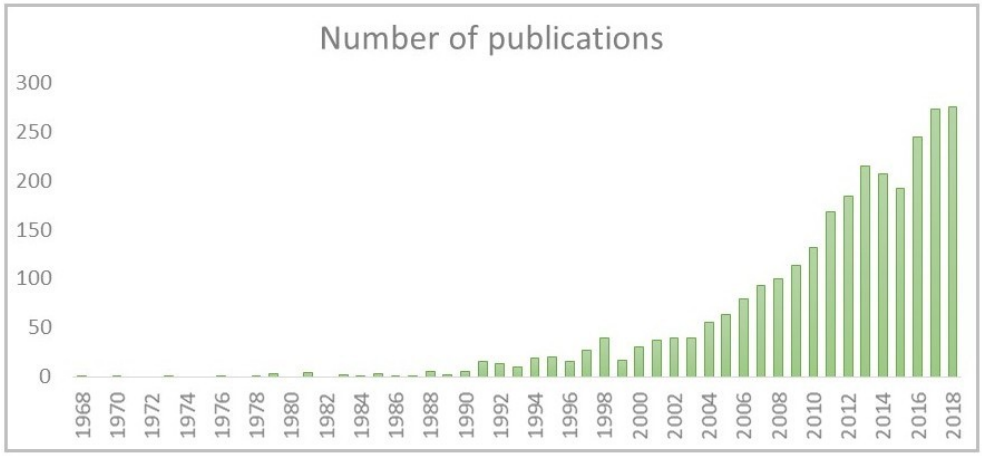


FIGURE 1.2: Number of publications with keyword "allostery". Data obtained from: www.webofknowledge.com [33].

on using only one protein structure which represents the lowest energy conformation. Normal modes are then calculated as eigenvectors of the Hessian matrix. Another increasingly popular methodology to study protein allostery is Elastic Network Modelling (ENM) [31, 32]. This method simplifies dynamics by representing the protein as an interconnected network of springs. Different levels of detail are possible, with particles either representing atoms, or groups of atoms. A model is then constructed around the most stable conformation and dynamics approximated as harmonic motion around this structure. Each of these methods have advantages and drawbacks, and combinations of techniques may be more appropriate, depending on the system under study.

These methods rely on allostery being explained by structural descriptors. However it is important to realise that allosteric effects may arise from very subtle changes in conformation or even dynamics that may be difficult to observe structurally, however may be validated by considering changes in energetics. Slight shifts in distributions of interaction energies or torsions could be deciding factors on enzyme activation or inhibition, therefore sensitive methods that also evaluate energetics are required to capture these subtleties.

1.4 Models to describe allostery

Initial descriptions of allostery were based on the haemoglobin oligomer, as structural details were available for this system, and described allostery in terms of cooperativity between adjacent chains. At this point, two mechanisms were predominant in explaining protein allostery which were briefly discussed in section 1.1, and while both are now known to be inadequate descriptions, they provided initial key insights into a complicated process, and could describe some of what was seen experimentally. Both were proposed in the mid-1960s; over a decade before molecular dynamics had been used to simulate proteins [34, 35], and so were at a point where atomistic descriptions of the dynamics of biomolecules were not possible. Nonetheless, they allowed significant insights into this complex process, and current models are expansions of these ideas. The Monod-Wyman-Changeux [2] (MWC) model describes allostery as a concerted process (figure 1.1), where the protein is either one of two states, an inactive and active form (termed tensed T, and relaxed R). Inter-conversion between the two occurs without the need for ligand binding and the population of each state is dependent on a thermal equilibrium. Binding of a ligand to either state then shifts this equilibrium in one direction or another, depending on whether the ligand is an activator or inhibitor. The result is that the allosteric effect is dependent not only on the concentration of bound ligand, but also on the equilibrium between the two states. Around the same time, Koshland, Némethy and Filmer [4] proposed another model (KNF model, which proposes that the active form, exists only in presence of the ligand via an "induced fit" mechanism. Activation of one subunit or binding site would then affect the binding of other subunits/binding sites, either by increasing or decreasing the affinity of the others. This explained something that the MWC model could not; negative cooperativity. Between the two models there is both agreement and conflict. Both models suggest that the protein can exist in two particular states, however one implies that these two states already exist in equilibrium, while the other suggests that the activated state only exists after ligand binding. Both models imply that allosteric proteins consist of more than one protein chain, although it is now known that even single

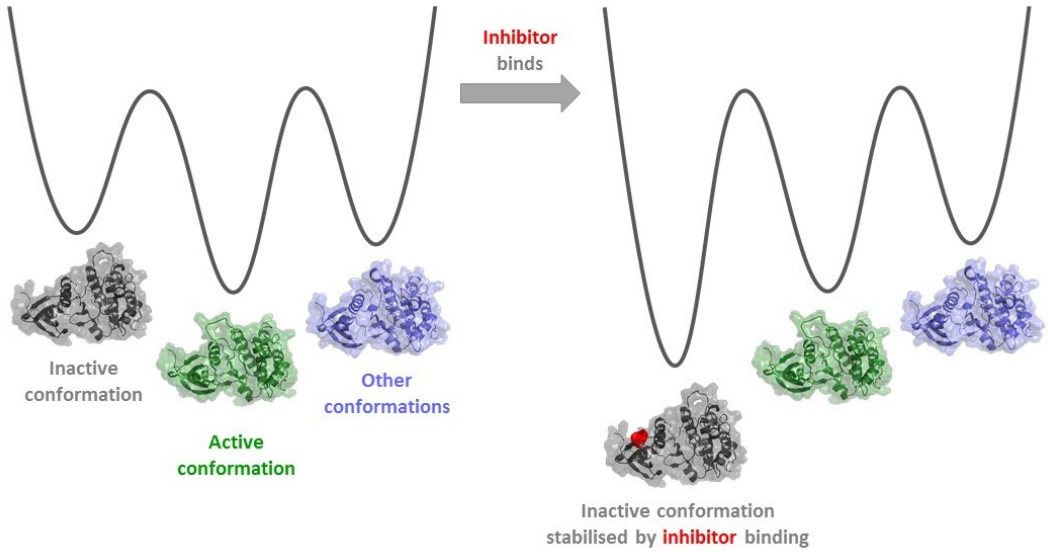


FIGURE 1.3: Conformational selection model for allostery. Many states exist, and binding of allosteric inhibitor stabilises inactive conformation.

monomeric proteins can undergo allosteric modulation.

The development of new experimental techniques [36] and molecular dynamics simulations [34, 35] changed the understanding of many aspects of protein dynamics and allosteric regulation. A recent description of allostery is based on conformational ensemble theory [37]. According to this model, proteins exist in a huge range of conformational states around their native state [38, 39], and the distribution of these states defines the conformational ensemble. Allosteric effects are a result of shifting of the distribution of these states, which occurs due to changes in the stability of a particular state caused by binding of an allosteric ligand (Figure 1.3). It is possible to stabilise different states, for example inhibitors will stabilise the inactive conformation, and activators the active conformation.

It is important to highlight that no new states are created according to this theory, but only a shift in the distribution of already existing states [40]. For example, binding of an activating ligand would stabilise a particular conformation of the protein which leads to products, and so move the equilibrium in this direction. Whereas, binding of an inhibitor would stabilise

the inactive conformer, and so reduce activity. Conversely, a structural view of allostery suggests that some conformational change occurs on binding of a ligand. This results in some form of correlated motions, which transmit a "signal" from the allosteric site to the orthosteric site, leading to changes in reactivity. In this model, it is not simply stabilisation of an existing conformer, but instead is more in line with the "induced fit" explanation of enzyme activity. To understand the conformational changes associated with an allosteric effect, the overall dynamic changes in conformation must be considered, as static images of either end state of the protein do not help in any understanding of how this change propagates from one site to another in order to affect this equilibrium. Two models of signal propagation have been suggested [41]. The first involves a series of successive motions from one site to the other via one pathway, which has been termed the "domino" model. The other model (the "violin" model) describes the signal via many smaller pathways, with changes in vibrational patters being responsible for the allosteric response (figure 1.4).

Although it is often the case, allosteric events do not necessarily imply vast conformational changes. In fact, correlated changes of a few torsional angles can result in very similar conformations, but direct certain key groups in slightly different ways which promote catalysis. Similar to the suggestion of the "violin" model, others had noted that conformational change is not even required [42], and that allosteric responses may *"travel across the structure as an 'energy signal'"*, in some cases more related to changes in dynamics than conformational changes [21, 43] . This theory was developed in the mid-80s, and describes allosteric effects as a result of changes in thermal fluctuations. Although this theory was presented some time ago, it was only relatively recently [44] confirmed experimentally, in the case of the protein calmodulin. While the two models of conformational ensembles, or structural changes seem independent from one another, some studies have shown that they may both be simultaneously valid [40]. It is also possible that not all allosteric proteins will fall under the same sort of mechanism. In

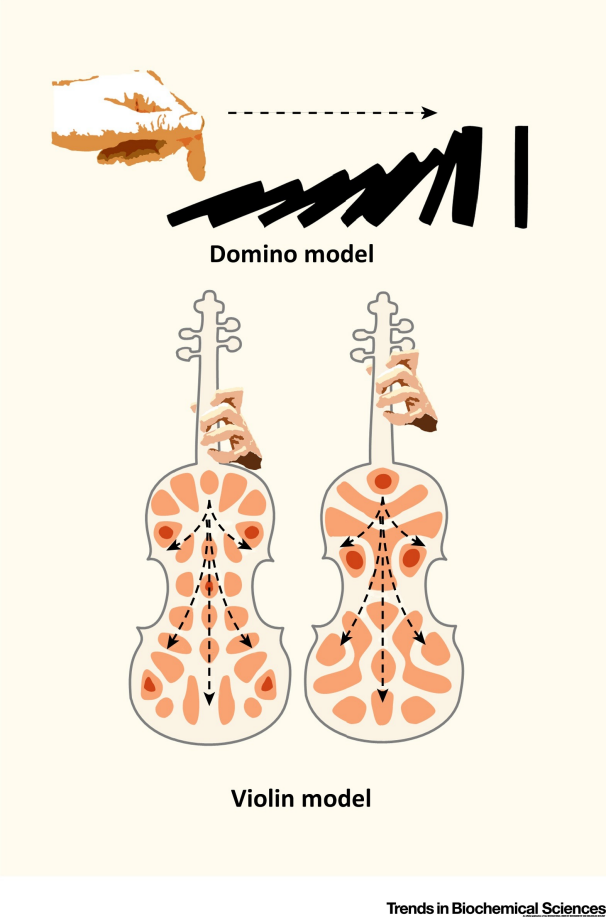


FIGURE 1.4: Figure taken from reference [41]. The "domino" model describes the allosteric effect as conformational changes which occur sequentially from one site to another. The "violin" model suggests that binding an allosteric ligand affects many regions of the protein, and is not a direct pathway. This could be structural or dynamic changes and affect multiple regions of the protein, which include the active site.

some cases, clear conformational changes occur, while in others allosteric effects can be seen with no significant conformational change. Therefore techniques that capture both the obvious and the subtle changes are required in order to fully understand allostery.

Chapter 2

Background theory

2.1 Statistical mechanics

In order to describe the dynamic properties of a system using MD simulations, there are several concepts from statistical mechanics which should first be defined.

In principle, to determine the macroscopic properties of a system, the solution of Newton's equations of motion over infinite time would give this result, if the system is ergodic. However, this is only useful for very small systems with only a few interacting particles, and is completely impossible for systems such as proteins with thousands of atoms. Therefore, we rely on computing distributions of configurations of the system using statistical mechanics, and using some assumptions, allow us to replicate the macroscopic properties of a large system of interacting molecules.

2.1.1 Ensembles

In statistical mechanics, an ensemble is a collection of different microscopic states of a system, whose average properties correspond to those observed for the system at the macroscopic level. Depending on the system under study, different ensembles may be of interest when running simulations (such as molecular dynamics). For example the canonical ensemble (NVT:

constant number of atoms, volume and temperature) or microcanonical ensemble (NVE: constant number of atoms, volume and energy). The isothermal–isobaric ensemble (NPT: constant number of atoms, pressure and temperature) is often used as it replicates the conditions under which chemical reactions are generally carried out experimentally.

2.1.2 Ensemble averages

Ensemble averages, as the name suggests, allow calculation of mean values of the observables of the system as a function of individual microstates; so while each microstate may have a different value for a particular property, the mean value will remain constant. The ensemble has a well defined property value that is postulated to correspond to the observable value in the limit of a large number of particles in the ensemble. In molecular dynamics, the assumption is that each one of these microstates will be visited proportionally to their Boltzmann probability over the length of the simulation. Therefore averaging over time should result in the same value as averaging over the large number of systems. This assumption, for a particular observable A, is that averages over time are equivalent to averages over ensembles is the Ergodic hypothesis:

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time} \tag{2.1}$$

In order to relate $\langle A \rangle_{ensemble}$ to $\langle A \rangle_{time}$, we must make several assumptions. First, it is important to note that individual microstates are not equally likely. Different microstates will contribute to a macrostate with varying weights, which depend on the energy of a particular microstate. Every instantaneous configuration (or snapshot of a MD simulation) of all molecules in the system represents a particular microstate. This configuration changes over time as populations of molecules in each configuration will change. The configuration of molecules in a particular microstate depends on the weight of that configuration, which is dependent on the energy of the state according to the Boltzmann distribution:

$$\frac{N_i}{N} = \frac{e^{-\beta\epsilon_i}}{\sum_i e^{-\beta\epsilon_i}} \tag{2.2}$$

where N is the total number of molecules, N_i is the number of molecules in state i, which have energy ϵ_i. The term β is a parameter which defines the temperature (T) for which we want to calculate the most likely population of each state, and the Boltzmann constant (k). It is defined as:

$$\beta = \frac{1}{kT} \tag{2.3}$$

The distribution of states therefore only depends on the energy of each state, and the ratio between any two given states (or the relative probability of a particular state) is defined by the Boltzmann factor:

$$\frac{p_i}{p_j} = e^{\frac{\epsilon_j - \epsilon_i}{kT}} \tag{2.4}$$

where p_i is the probability of state i, ϵ is the energy of a state, k is the Boltzmann constant, and T is the temperature of the system.

2.2 Simulation methods

2.2.1 Force fields

For systems such as proteins, which contain many atoms, it is not possible to use quantum mechanics to describe the energy of a microstate as the computation required would take too much time. Therefore methods which describe atoms in terms of nuclear motions rather than electronic motions must be employed, as this makes it possible to run calculations on very large systems with many atoms within a reasonable time frame.

The energy of the system is described in terms of a potential energy function and a kinetic energy function, along with a set of parameters which are obtained from experimental data or quantum mechanical calculations. The combination of the potential energy function and parameter set is called a force field, and many different force fields exist. The choice of force field used will depend on the type of system being studied, as parameters will be

optimised for particular cases, such as for proteins or for organic molecules. The function which is used, includes terms to describe intramolecular motions (with terms for bond lengths, angles and torsions), and intermolecular forces (with terms for van der Waals and electrostatic interactions). For example, in the AMBER force field [45] the functional has the form shown below:

$$U(r^N) = \sum_{bonds} \frac{1}{2} k_i (l_i - l_{i,0})^2 + \sum_{angles} \frac{1}{2} k_i (\theta_i - \theta_{i,0})^2$$

$$+ \sum_{torsions} \frac{1}{2} V_i (1 + cos(n\omega - \gamma)) \tag{2.5}$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma ij}{r_{ij}} \right)^1 2 - \left(\frac{\sigma ij}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)$$

The first summation is associated with stretching of covalent bonds, and is approximated by a simple harmonic equation, where k is the force constant, l_i is the bond length at a given time, and $l_{i,0}$ is the equilibrium bond length.

The second summation takes into account deviations from equilibrium bond angles, where again k_i is the force constant associated with this motion, θ_i is the angle and $\theta_{i,0}$ is the angle at equilibrium. The third summation is a function which describes energies involved in rotation about bonds. V_i is associated with relative barrier energies, with higher values for types of bond rotation which require more energy. The term n is the multiplicity [46], which describes how many minima in energy will be involved to complete a full rotation. γ describes where these minima are, and ω is the torsional angle. The last double summation in this function describes all pairwise non-bonded interactions, which defines van der Waals interactions using a Lennard-Jones potential and electrostatics with a Coulomb potential. In the Lennard-Jones equation, r_{ij} is the distance between two atoms i and j, ϵ describes the well depth and σ is the distance for r_{ij} which has energy equal to

zero. These are illustrated in figure 2.1. In the Coulomb potential, q_i and q_j are partial charges on atoms i and j, ϵ_0 is the vacuum permitivitty, and r_{ij} is the distance between two atoms i and j.

2.2.2 Long range interactions

In practice, it is too computationally expensive to calculate long range non-bonded interactions for every pair of atoms in the system. Usually, for the Lennard-Jones potential this is handled by truncation of the function at a certain cutoff distance, as shown in figure 2.1.

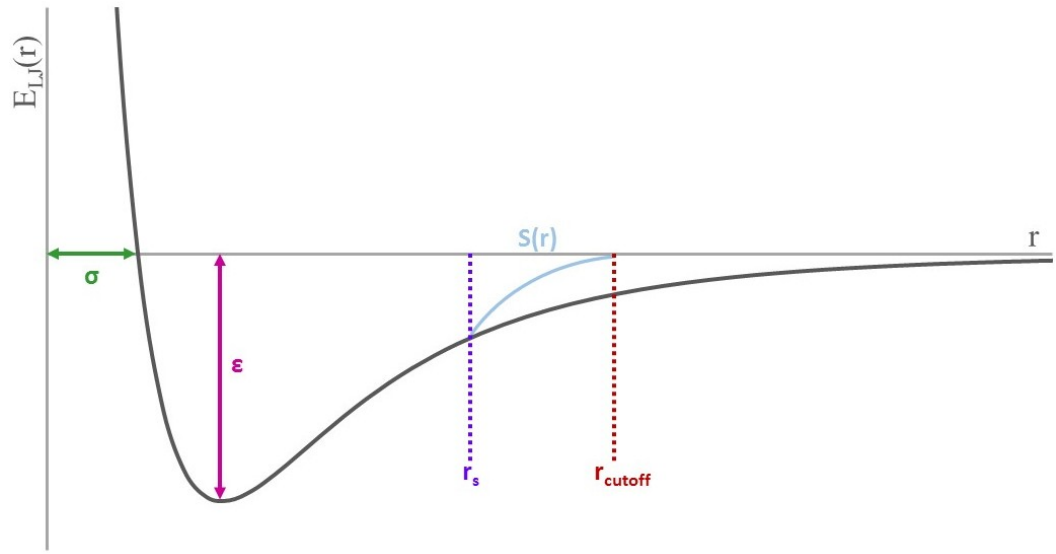


FIGURE 2.1: Lennard-Jones potential describing the non-bonding component of the potential energy arising from van der Waals interactions. A cutoff is applied at a distance r_{cutoff}, above which interactions do not contribute to the potential energy.

In order to avoid a discontinuous function, a switching function ($S(r)$) is often added to reduce the value of the function gradually to zero as it reaches the cutoff distance (r_{cutoff}), as detailed in equation 2.6.

$$E_{LJ} = \begin{cases} E_{LJ}(r) & \text{when } r \leq r_s, \\ E_{LJ}(r)S(r) & \text{when } r_s \leq r \leq r_{cutoff}, \\ 0 & \text{when } r > r_{cutoff} \end{cases} \tag{2.6}$$

However treatment of the Coulombic term at long distances is less straightforward, as electrostatic interactions can occur at longer distances than those from dispersion. Different methods can be applied such as PME (particle mesh Ewald), or the Reaction Field method.

Using the PME method [47], electrostatics are treated differently depending on either they occur at longer or shorter distances. An Ewald summation involves the treatment of longer range distances to be calculated using a summation in Fourier space which is possible due to periodicity, and the short range interactions are calculated with direct summation in real space. PME is a method which allows the approximation of this Ewald sum to be done numerically [48].

$$E_{Elec(Ewald)}(r) = E_{short}(r) + E_{long}(r) \tag{2.7}$$

Where E_{short} are the short distance interactions, and E_{long} are the long distance interactions, at a distance r. These can be defined as:

$$E_{short}(r) = \sum_{i,j} E_{short}(r_j - r_i) \tag{2.8}$$

and:

$$E_{long}(r) = \sum_{k} \tilde{U}_{long}(k)|\tilde{p}(k)|^2 \tag{2.9}$$

where $\tilde{U}$ is the Fourier transform of the potential, and p(k) is the Fourier transform of the charge density.

In the reaction field method [49, 50], a cutoff (r_{cutoff}) is applied to each atom in the system, which allows for all interactions within the cutoff to be treated explicitly, and outwith the cutoff there is a medium with uniform dielectric constant (ϵ_{RF}), which can be polarised by interaction with molecules

within the cavity. The electric (reaction) field (E_a) that this produces is described by:

$$E_a = \frac{2(\epsilon_{RF} - 1)}{2\epsilon_{RF} + 1} \frac{1}{r_{cutoff}} \sum_b \mu_b \tag{2.10}$$

where $\sum_b \mu_b$ is the summation of dipole moments for molecules within the cavity. The potential then becomes:

$$E_{Elec(RF)} = E_{Coulomb} + E_{RF} \tag{2.11}$$

Where $E_{Coulomb}$ is the term shown for the Coulomb potential in equation 2.5, and E_{RF} is the component from the reaction field. This is defined as:

$$E_{Elec(RF)} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i} q_i q_j \left[\frac{1}{r} \left(1 + \frac{\epsilon_{RF} - 1}{2\epsilon_{RF} + 1} \left(\frac{r_{ij}}{r_{cutoff}} \right)^3 \right) \right] \tag{2.12}$$

2.2.3 Molecular dynamics

The computational method of molecular dynamics [35] is based on a numerical and sequential solution to the classical equations of motion (equation 2.13) in order to obtain information on the macroscopic properties of the system. Starting with an initial arrangement of atoms, and by randomly assigning velocities based on the Maxwell-Boltzmann distribution, forces can be calculated by differentiating the energy function which describes interactions with neighbouring atoms. The acceleration is then computed from Newton's second law, and the particle is then moved to a new position. This is then repeated for a series of timesteps, and during each step the force remains constant. Hence the calculation must be broken down into a series of many, very short, timesteps.

$$\vec{F} = m\vec{a} = m\frac{d\vec{v}}{dt} = m\frac{d^2\vec{r}}{dt} = -\nabla U(\vec{r}) \tag{2.13}$$

In equation 2.13, $\vec{F}$ is the force of a particle, m and $\vec{a}$ are mass and acceleration respectively, $\vec{r}$ is the position vector, and $U(r)$ is potential energy

function, described by a force field. Integration of each position and velocity at each time step defines a trajectory, from which a wide range of structural and dynamical information can be obtained, as well as thermodynamic properties.

2.2.3.1 Integration methods

In order to propagate the dynamics of the system, numerical methods to solve the equations of motions are required. Common integration algorithms are variations of Verlet integration. An initial configuration of the system (usually obtained from a X-ray crystallographic structure), defines positions for each particle at $t = 0$. The algorithm then considers that the positions of the particles in the system can be approximated by a Taylor series:

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + ... \tag{2.14}$$

where r is the position of the particle, v is the velocity, a is the acceleration, t is the time, and δt is the timestep. Positions for time $t - \delta t$ can then be defined as:

$$r(t + \delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \tag{2.15}$$

Therefore to define the new position at time $t + \delta t$, the summation of equations 2.14 and 2.15 can be combined, to give:

$$r(t + \delta t) + r(t - \delta t) = 2r(t) - r(t - \delta t) + a(t)\delta t^2 \tag{2.16}$$

The Velocity Verlet algorithm [51] is an extension of the above, which in addition to position also calculates velocity at each δt. The Leap-frog algorithm is similar to the Velocity Verlet, however instead of calculating both directly at time $(t + \delta t)$, velocities are calculated first at time $(t + \frac{1}{2}\delta t$, followed by calculation of positions at time $(t + \delta t)$. This results in velocities being calculated at different time to the positions, which can be corrected using the following equation:

$$v(t) = \frac{1}{2}\left[v(t - \frac{1}{2}\delta t) + v(t + \frac{1}{2}\delta t)\right] \tag{2.17}$$

The timestep δt is selected such that sufficient conformational space can be sampled, but also provide numerical stability. If δt is longer than the timescale of the fastest motion, the energy of the system will not be conserved.

2.2.3.2 Periodic boundary conditions

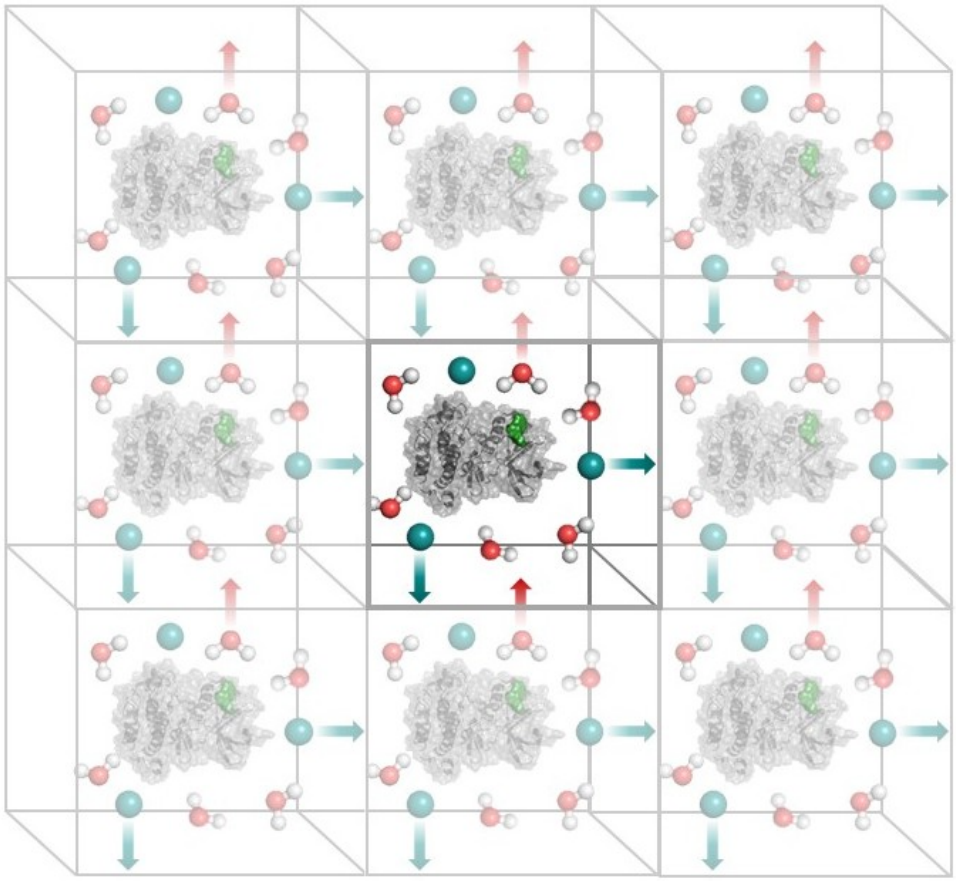


FIGURE 2.2: Periodic boundary conditions.

There is a limit to the number of particles which it is possible to simulate using molecular dynamics at a reasonable computational cost. Boundaries must be carefully considered, otherwise the ratio of surface molecules to

bulk will be too high, and surface effects become an issue. Periodic boundary conditions allow the particles which are near the edges of the box to be treated as if they were in the bulk. As a particle moves out of a particular box, it is replaced by an identical particle that appears from the opposite side of the box. Therefore the number of particles within each box remains constant (Figure 2.2).

2.2.3.3 Thermostats and barostats

Depending on which thermodynamic ensemble molecular dynamics is being run under, pressure and/or temperature may need to be fixed around a certain value during the course of the simulation. Various methods exist to achieve this, and ensure that average values of pressure or temperature remain constant.

Thermostats

It is important to control the temperature of the system during an MD simulation, which is calculated from the kinetic energy using the equipartition theorem. The instantaneous temperature of the system is likely to vary from the temperature selected to run the simulation, however by averaging over many instantaneous temperatures, the target temperature should be achieved. Thermostats are required to ensure that these time averaged values are constant, and ensure that the correct ensemble is sampled. Different algorithms have been designed in order to adjust the solution to the equations of motion in order to maintain an average temperature required for the system under study.

The Berendsen thermostat [52] is a weak coupling algorithm, which using a scaling factor to alter the momenta of molecules to that of the temperature set for the simulation. This ensures that the kinetic energy of the system is that of the required temperature. However there are some issues which can result from use of this thermostat, which leads to an uneven distribution of energy between the different degrees of freedom, and therefore

does not maintain the equipartition theorem that energy is divided equally by all degrees of freedom.

The Andersen thermostat [53] adjusts temperature by selecting a particular molecule and adjusting the temperature of that particle by altering the velocity from the Maxwell-Boltzmann distribution for the temperature defined for the simulation. This can cause artefacts in the dynamics, as it relies on sudden changes in velocity. Other thermostats such as Nosé-Hoover [54, 55], or the Langevin thermostat may improve on some of these issues, however each have individual benefits and drawbacks.

Within this work, the Andersen thermostat was used, as it is the only thermostat implemented in the software used to run the majority of MD simulations, SOMD (Sire/OpenMM) [56]. However, care was taken to check for issues that can occur with use of this thermostat, by monitoring backbone RMSD.

Barostats

In order to replicate experimental conditions, it is often desirable to simulate the system under the NPT (or isothermal-isobaric) ensemble. In order to maintain a constant average pressure, a barostat algorithm is used. The Berendsen barostat [52] works in a similar way to the Berendsen thermostat, in that the volume is altered by some scaling factor by addition of terms to the equations of motion, to allow fluctuations in the pressure to remain around the target pressure. The Monte-Carlo barostat adjusts the pressure by assessing both the positions and the volume of the system at alternating timestep. The movement of a particle is either accepted or rejected based on the Metropolis criterion. To alter the volume of the system, a random expansion or compression is applied, and either accepted or rejected depending on the ratio of energies before and after the adjustment. Other barostats such as Andersen [53], Nosé-Hoover [55], and Parrinello-Rahman [57] are commonly used.

In all simulations run with the software SOMD (Sire/Open MM) [56], the Monte-Carlo barostat was used.

2.3 Information theory

2.3.1 Kullback-Leibler divergence

Often, methods to analyse molecular dynamics trajectories involve analysis of probability distributions of some parameter obtained from the the MD simulation. There are many statistical methods to determine differences between these distributions of data to gain useful information. F-divergences are functions which provide information on the distance between two probability distributions. One such function is the Kullback-Leibler [58] divergence (KL). Starting from a reference distribution (e.g. obtained from a molecular dynamics simulation), the distance is measured to reach a target distribution (from another simulation, of the same or different system). For example, by taking a reference distribution from a simulation containing a protein only and the target distribution from a simulation of a protein and ligand complex, the divergence of the target from the reference distribution can be obtained. This essentially gives information relating to the extent of differences between probability distributions, otherwise known as the relative entropy. Calculation of the KL divergence of a particular observable can allow quantitative comparisons of different systems. The KL divergence is defined for a discrete distribution in equation 2.18, where P and Q are the probability distributions from the target and reference ensemble respectively, N is the number of bins, and $P(i)$ or $Q(i)$ are the counts of the i-th bin in distribution P or Q.

$$D_{KL}(P \parallel Q) = \sum_{i}^{Nbins} P(i) \ln \frac{P(i)}{Q(i)} \tag{2.18}$$

This is taken to mean the distance of distribution Q from distribution P, and is described as such due to the asymmetry of the relationship, in that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. The value of the distance between two distributions is always greater than, or equal to zero, with a value of zero only if the two distributions are identical.

Even if considering two separate molecular dynamics simulations of the same system, it is likely that the KL value will not be zero as some statistical

variation is expected, due to finite sampling. If the term $\frac{P(i)}{Q(i)}$ in equation 2.18 is replaced only by $P(i)$, then this equation is a measure of the entropy of a distribution, however by taking this ratio, a difference between the two distributions is defined. In the case where distributions P and Q are completely continuous, equation 2.18 gives a solution. An issue arises when comparing probability distributions which are not continuous, as either $P(i)$ or $Q(i)$ is equal to zero, leading to an undefined divergence. One approach to resolve this issue is to add a prior count to all histogram bins. The simplest prior adds a uniform amount to each bin as described in equation 2.19 to give a new distribution P'.

$$P'(i) = P(i) + x\overline{P} \tag{2.19}$$

In this equation, $P(i)$ represents bin number i from the original distribution, x represents a fraction which influences the size of the value added and $\overline{P}$ represents the mean of the original distribution P. Therefore all bins contain a non-zero value, and assuming this prior count is significantly smaller than the probabilities obtained from the data, the KL value obtained should still be representative of the distance between those two distributions. However, a problem arises using this method when the differences in overlap vary considerably between systems. For those which have only a small region which is non-overlapping, the effect will be different to those with a larger non-overlapping region. Therefore comparison of these two cases would be difficult. To resolve this issue, a fixed value can be split over only the bins which have a zero count, therefore adding differing values per bin depending on the number of empty bins. This should allow weighting to compensate for the differing number of empty bins. Methods to determine how many bins (N) to use to plot a histogram vary considerably, and there is no general rule which will suit all data. One commonly suggested method is to use a number of bins equal to the square root of the number of data points. This gives a good initial estimate for an appropriate number of bins, however as no one method suits all data, a range of bin numbers must be tried, in order to determine how this affects the resulting distribution. Testing must be carried out to establish how the number of bins used affects the

KL value obtained, and on the value used for x. An example of such tests can be found in section 3.3.3.1.

2.3.2 Jensen-Shannon divergence

In some cases, it may be more useful to have divergence values which are symmetric to both variables P and Q. While KL divergence is acceptable to use when comparing all to the same reference (i.e. the inhibitor bound simulation is always Q (reference), while P is any activating ligand bound simulation), it is less useful if we want to compare many different compounds to each other as we cannot always select the same reference. In this case, we compute the Jensen-Shannon divergence which is symmetric.

$$\mathrm{JSD}(\mathrm{P} \parallel \mathrm{Q}) = \frac{1}{2}\mathrm{D}(\mathrm{P} \parallel \mathrm{M}) + \frac{1}{2}\mathrm{D}(\mathrm{Q} \parallel \mathrm{M}) \tag{2.20}$$

2.3.3 Mutual information

Various structural differences may be observed between different ligand bound simulations, however determining which are responsible for differences in activity is challenging. In order to describe this, Mutual Information (MI) can be used to show correlations between different variables. MI is able to capture higher order correlations where small changes in one parameter cause larger changes in another. Taken from information theory, MI describes the amount of information which one variable contains about another variable. MI can take any value from 0 to $+\infty$, however normalisation is commonly done to yield values between 0 and 1 (where 1 is completely correlated and 0 means completely uncorrelated).

To mathematically define MI, we should first introduce the concept of information entropy. Analogous to the equation which describes entropy in statistical mechanics, information entropy is defined as:

$$H(X) = -\sum_{x_i}^{n} P(x_i) \log_b P(x_i) \tag{2.21}$$

This defines the entropy for a variable X, which can have values belonging to $(X) = \{x_1, x_2, x_3, .., x_n\}$. The units of entropy depend on the base on the logarithm, where base 2 results in units of "bits", base e gives units of "nats", and base 10 is units of "bans".

Information entropy defines how much information is contained in a set of one variable. Mutual information then extends this, to define how much information we can obtain from one variable, by observation of another distinct variable. Therefore we utilise this in molecular dynamics simulations, in order to define how correlated two different variables are.

The MI of two discrete variables X and Y is defined as:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2.22}$$

Where X and Y are two different variables, where variables X and Y can have values $\{x_1, x_2, x_3, .., x_n\}$ and $\{y_1, y_2, y_3, .., y_n\}$ respectively, and probability distributions of $P(X)$ and $P(X)$.

The relationship of MI to information entropy can be expanded by considering equations for conditional entropy:

$$H(X|Y) = -\sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(y)} \right) \tag{2.23}$$

And joint entropy:

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log \left(p(x,y) \right) \tag{2.24}$$

From here we can define MI as:

$$I(X;Y) = H(X) - H(X|Y) \tag{2.25}$$

Or as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.26}$$

In practice, because distributions are obtained from finite samples, it is possible to observe artefactual correlations between independent variables using MI. To account for this, a correction was made to subtract a value associated with noise. This was achieved by computing MI with one of the two variables randomised in time (X_{rand}). Therefore, MI reported in the results is as follows:

$$I_{corr}(X;Y) = I(X;Y) - I(X_{rand} : Y) \tag{2.27}$$

2.3.4 Principal component analysis (PCA)

In order to extract information from MD trajectory, motions which are important for biological function must be identified. Considering only Cα atoms, a protein simulation with 286 residues (as for PDK1) would have 858 dimensions. Finding which motions are functionally important in this high dimensionality space is not trivial. Methods which highlight major collective motions are needed, to reduce the dimensionality into a few important modes (principal components). In most PCA studies of protein dynamics, the implicit assumption is that the dominant collective modes found by this technique are the main functional modes, i.e. larger variance components are dynamically interesting, while others are 'thermal noise'. This is because it is often the case that the functionally relevant motions of a protein involve displacement of many atoms at the same time, and so Cartesian coordinate PCA can be used to highlight these functionally important modes. To use PCA to study smaller, yet functionally relevant local motions, it is possible to select a subset of atoms/residues and perform PCA over this subset of atoms, thus highlighting smaller amplitude motions which could be important. This approach could be useful for applying PCA to understanding allostery, as functionally relevant differences in structures are likely to be located between active and allosteric sites, and do not always involve large differences in overall structure.

PCA is a multivariate statistical method which allows reduction of the dimensionality of a molecular dynamics trajectory by projecting a set of N-dimensional data onto a new coordinate space, where the first few new dimensions explain the largest variance in the dataset. This makes it possible to describe a large percentage of the overall variance in the data using only a few principal components (dimensions) rather than the original N dimensions of the original dataset.

For a set of data X, PCA aims to construct a new set of variables Y, such that:

$$Y = W'X \tag{2.28}$$

Where W' is a term obtained from the PCA, which results in the variables of Y being a weighted average of the original set of variables. W is obtained from the covariance matrix, as described below. Covariance is defined as:

$$Cov(A, B) = \frac{\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})}{n} \tag{2.29}$$

where n is the number of degrees of freedom, A_i and B_i are values for two different degrees of freedom, and $\overline{A}$ and $\overline{B}$ are their mean values. Positive covariance means that both A_i and B_i are correlated, while a negative covariance means A_i and B_i are anti-correlated. A value of zero covariance means that A_i and B_i are completely uncorrelated.

A covariance matrix is then constructed with the format:

$$\begin{vmatrix} Variance(x_1) & Covariance(x_1, x_2) & ... & Covariance(x_1, x_n) \\ Covariance(x_2, x_1) & Variance(x_2) & ... & Covariance(x_2, x_n) \\ ... & ... & ... & ... \\ Covariance(x_n, x_1) & Covariance(x_n, x_2) & ... & Variance(x_n) \end{vmatrix}$$

Diagonalisation of the covariance matrix yields a set of eigenvectors (principal components), each with an associated eigenvalue. The eigenvectors are ranked by eigenvalue, and those with the highest eigenvalues are the

principal components. A set of eigenvectors which have the largest eigenvalues are selected, and form a mxn dimensional matrix W. W' which is shown in equation 2.28 is then the transpose of W.

From a MD trajectory, a covariance matrix is constructed for a particular descriptor of interest. In this work, Cα coordinates have been used (264 Cα for the PDK1 system), and so the covariance matrix will be 792 by 792 (as x, y, z for each Cα) however other metrics such as torsional angles could be used [59].

2.4 Energy decomposition

As was mentioned earlier, the functional form of the potential energy surface is defined in equation 2.5. This is pairwise additive over all atom pairs i,j, and so can be decomposed for every residue pair (I,J), by summing every i,j atom pair interaction for atoms belonging to each residue.

From this, we can therefore obtain the contribution to $U(r)$ which relates directly to the non-bonding interactions by considering the terms within the last double summation, which relate to the van der Waals interactions, and electrostatic interactions. We can then decompose this potential energy component into a pairwise value, which relates specifically to the interaction of one residue in the system (a ligand, molecule, or protein amino acid) to any other residue in the system.

2.5 Markov state models

A Markovian process defines a stochastic series of events, where probabilities of future events are determined only by the current state, and not by any preceding events. A Markov model is constructed of a set of discrete states, which are connected by the probabilities from moving from one state to another, referred to as transition probabilities.

The goal of building a Markov model from MD simulation data is to transform the time series data obtained from MD trajectories into a series

of discrete states, and then determining the transition probabilities between each state in order to gain mechanistic insight.

To build a Markov mode, the following process is followed:

- Reduce dimensionality.
- Cluster to obtain microstates and assign structures to each microstate.
- Assign microstates to slowly converting macrostates.
- Estimate populations of each macrostate and transition probabilities between each macrostate.

Clustering of the initial simulation data is used to construct a series of microstates. The trajectories are then converted from a series of coordinates over time, to a series of states over time. A lag time (τ) is defined as the timestep for each transition from state i to state j. For a given lag time τ, a transition matrix $P(\tau)$ is then constructed, such that:

$$P(\tau) = [p_{ij}(\tau)] \tag{2.30}$$

In this equation, p_{ij} is the conditional probability of the transition from state i to state j at a given lag time τ. The value for τ is selected such that the processes of interest are well resolved, but also allows the model to obey Markovianity. The implied timescales (ITS) for each process are determined by the following equation:

$$t_i = \frac{-\tau}{ln|\lambda_i(\tau)|} \tag{2.31}$$

A timescale is selected which allows for the smallest value of τ for which the ITS converge to a constant value, which is no longer dependent on the value of τ. In equation 2.31, t_i is the ITS of the i^{th} process, λ_i are the eigenvalues of $P(\tau)$, and τ is the lag time. To confirm that the processes described by $P(\tau)$ are Markovian, the Chapman Kolmogorov (CK) test [60] can be carried out, which determines whether $P(\tau)$ obeys the CK property:

$$P(k\tau) = P^k(\tau) \tag{2.32}$$

$P(k\tau)$ is a MSM constructed by multiplying the lag time τ by a factor k, and $P^k(\tau)$ is the MSM built at lag time τ to the power of k. The eigenvalues and eigenvectors of the transition matrix $P(\tau)$ are then calculated. The stationary distribution (π) is given by the left eigenvector associated with the highest eigenvalue (which is equal to 1) $\lambda_1(\tau)$:

$$\pi^\top P(\tau) = \pi^\top \tag{2.33}$$

Further eigenvectors provide information on the dynamics, and coefficients of each eigenvector (λ_2,....,λ_i) describe transitions into and out of each state.

As the MSM is constructed using conditional probabilities between different states, it is not not required to comprehensively sample the process of interest. This allows for a series of parallel MD simulations to be run in a much shorter time [61] than if the same processes were sampled using long equilibrium MD simulations.

2.6 Enhanced sampling methods

It is often the case that the particular motion of interest in a protein is not sampled within the normal equilibrium MD timescales (in the order of μs) and as it is not required to sample the full equilibrium distribution in order to construct an MSM, shorter simulations started from a range of different starting points can be used. In order to generate starting points for intermediate structures, different enhanced sampling methods can be used. In this work, steered MD (sMD) [62] has been used.

It is possible to modify the Hamiltonian of the system in order to bias the motion of a particular set of atoms or residues, to pull one region of the protein along a particular collective variable, S.

$$\begin{aligned} H_\lambda(X,t) &= H(X) + U_\lambda(X,t) \\ &= H(X) + \frac{k(t)}{2}(S(r(t)) - S(r_0) - \nu t)^2 \end{aligned} \tag{2.34}$$

In equation 2.34, S_0 and S are the initial and current values of the collective variable, v is the pulling speed and k is the force constant. Different collective variables are possible, and combinations of different distances can be used. In this work, the RMSD relative to a particular conformation was used.

Chapter 3

Allosteric modulation of phosphoinositide-dependent kinase-1 (PDK1) mediated by covalently bound small molecules

3.1 Introduction

3.1.1 Protein kinases as a drug target

Protein kinases are a class of allosteric enzyme, which regulate activity of their substrate proteins via phosphorylation [63]. The activity of the kinase itself can be modulated allosterically, as they have binding sites which are distant to the active site, which control the on/off switching of phosphorylation events. The functional effect of this phosphorylation on the substrate protein is also often allosteric, in that phosphorylation of serine, threonine or tyrosine at one site of an enzyme or receptor initiates a response at another site of that substrate protein. Kinases have been extensively studied, as they have been implicated in many different disease pathways, including cancer, diabetes, and many others [64].

The catalytic domain of protein kinases comprises of an N and C terminal lobe, with a binding site for ATP (the orthosteric, or active site) located between the two lobes. Coordinated to ATP either one or two Mg^{2+} ions, depending on the particular kinase. The important structural features

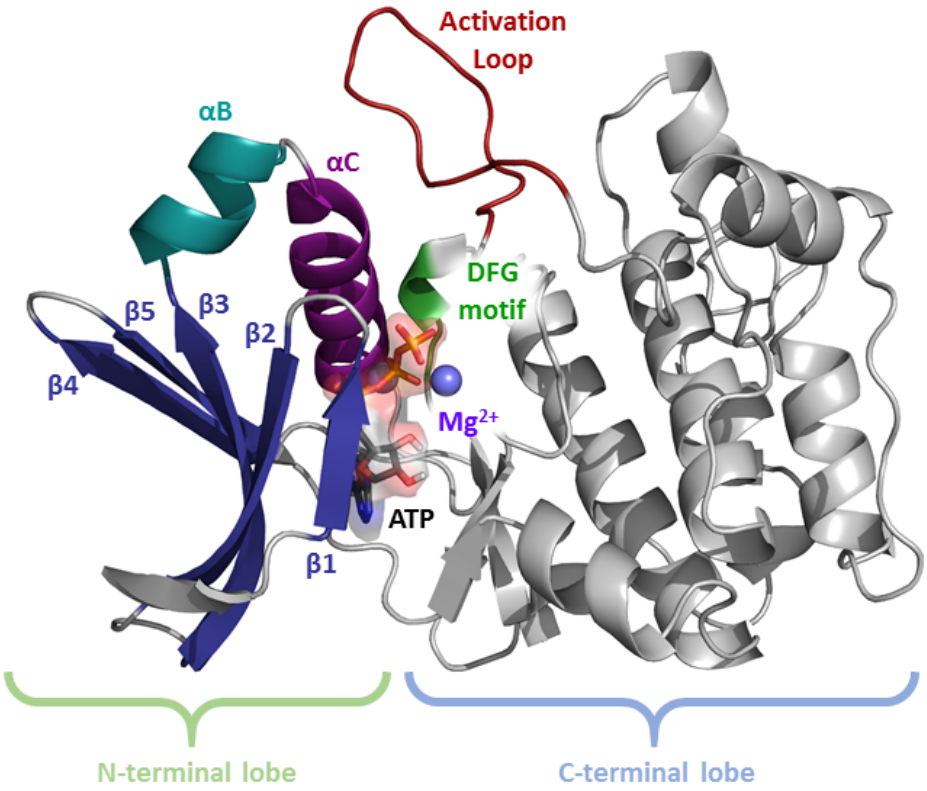


FIGURE 3.1: Common structural features of protein kinases, illustrated using the structure of PDK1.

which can be found in all kinases are highlighted in figure 3.1, which is illustrated using the protein kinase PDK1. In addition, most kinases have a sequence which extends from the C-terminal lobe termed the hydrophobic motif (HM), which usually contains a phosphorylation site [65, 66]. The structure of the HM is illustrated later in figure 3.7.

The human "kinome" (the set of all protein kinases which are encoded by the human genome) comprises of at least 518 different protein kinases, making them one of the largest families of enzymes and representing around 1.7% of human genes [67]. However this 1.7% can actually affect far more of the proteome: up to 30% of all human proteins [68] are phosphorylated by kinases. They are involved in the regulation of almost all important cell processes, including protein synthesis, transcription, cell division and growth,

and apoptosis. The 518 kinases are classified depending on which residues they phosphorylate. Most of these are serine/threonine kinases (STKs), followed by tyrosine kinases (TKs), and the remainder are capable of phosphorylating all of the above (DSKs - dual specificity kinases). The next classification is to split these into two groups: the majority (478), which have a eukaryotic kinase domain (ePKs - eukaryotic protein kinases); and around 40 which do not (aPKs - atypical protein kinases) [67, 69]. Of the ePKs, there are then 8 subfamilies, which are classified based on sequence similarity:

- CAMK (Ca^{2+} / calmodulin-dependent kinases)
- TKs (tyrosine kinases)
- TKLs (tyrosine kinase-like)
- CMGC (related to CDKs (cyclin-dependent kinases), MAP kinases (mitogen-activated protein kinases), GSKs (glycogen synthase kinases) and CDK-like kinases)
- STE kinases
- CK1 (casein kinase 1)
- AGC (related to protein kinase A, protein kinase G and protein kinase C)
- RGC (receptor guanylyl cyclase)

Both phosphorylation and dephosphorylation are required to achieve regulation. Phosphorylation of a protein or enzyme can affect the conformation or dynamics, change the enzyme's activity, or alter protein-protein interactions. Often phosphorylation results in activation or inhibition of an enzyme or receptor. This means that phosphorylation and dephosphorylation can act as a switch within complex signalling pathways, with protein kinases carrying out the phosphorylation, and protein phosphatases the dephosphorylation. Signal transduction occurs by successive activation and inhibition. Each of these steps must be tightly controlled, and dysregulation at any one of these steps is often a contributing factor to many diseases.

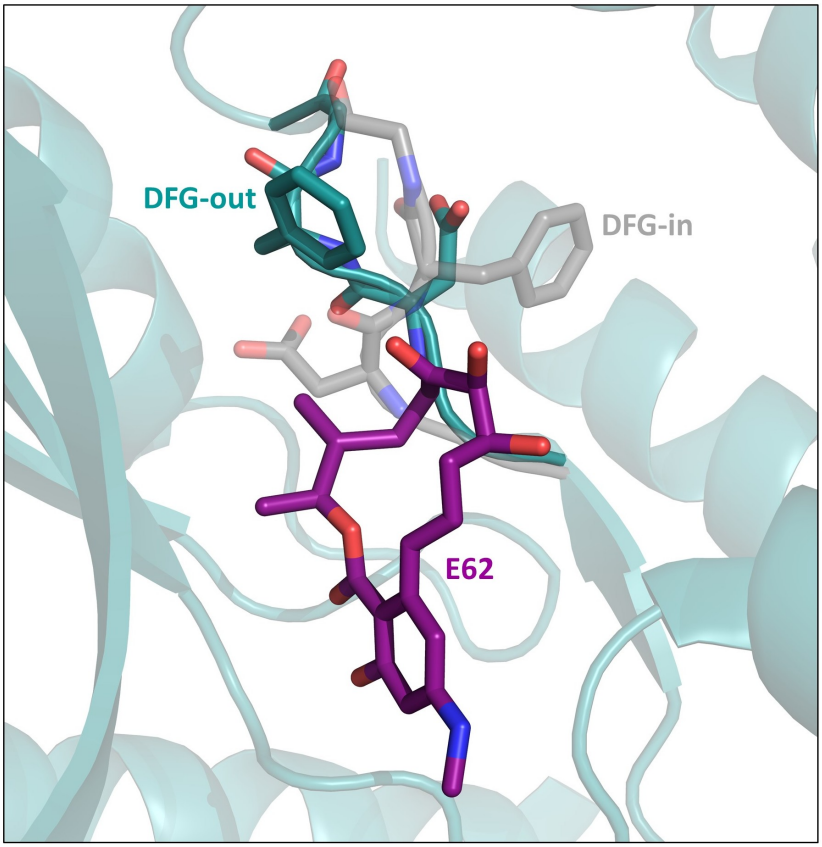


FIGURE 3.2: MEK1 kinase inhibitor E62 bound to the DFG-out conformation from PDB ID 5HZE. The DFG-in conformation of MEK1 is shown in grey (PDB ID 3W8Q).

Overexpression can lead to an over-abundance of a particular kinase, or mutation can lead to a kinase which is incorrectly in a permanently active state, leading to over-phosphorylation of its substrates. One broken link, and an entire signalling pathway can be led astray, therefore the development of approaches to *selectively* and accurately modulate activity are crucial.

As a result, kinases have been a focus for drug design for many years, however the interest has historically been in active site compounds. Depending on the binding mode, there are two classifications of active site compounds, type I or type II. The difference depends primarily on the conformation of the DFG motif, where type I binds the DFG-in, and type II the DFG-out (Figure 3.2).

These are both ATP competitive, however type II inhibitors can occupy

both the active site and an adjacent pocket. There are both type I and II kinase inhibitors in use, however few are particularly selective. Type I inhibitors such as Gefitinib [70] or Bosutinib [71] bind directly at the ATP site, and usually do so with some form of adenine analogue. Some include extensions which interact with other nearby sites, to attempt to promote selectivity. However due to the high conservation of the ATP site, selectivity is often poor and so other kinases are also inhibited, leading to adverse side effects.

Type II inhibitors bind to the ATP site, however as they bind to the DFG-out conformation, they have access to bind also to a pocket available only in this conformation (figure 3.3). Some successes of selective active site compounds have been for type II active site inhibitors for particular kinases, such as Imatinib (otherwise known as Gleevec) [72]. Imatinib is an inhibitor of several tyrosine kinases, and does so with reasonable specificity for Bcr-Abl tyrosine kinases, which are expressed as constitutively active due to a chromosome defect. Computational studies to determine the reasons behind the specificity suggest that conformational selection plays a crucial role. The findings suggest that Abl kinase has a preference for DFG-out conformation even without ligand binding [73], which is the favoured conformation for binding of imatinib. However, in related kinases the DFG-in conformation is preferred. While this has been an impressive development, the mechanism of the selectivity was not discovered until much later, and rational design of selective active site inhibitors is still extremely challenging. In addition, it may only be some specific kinases for which this is possible, as this conformational selectivity around the ATP site may only be useful in select cases.

A further class of inhibitors of kinases are allosteric, or type III inhibitors. This subset of molecules has been defined as those which bind directly *next* to the ATP site, however do not compete with ATP binding (figure 3.4). The first type III allosteric small molecule kinase inhibitor to be developed was Trametinib, a MEK1/MEK2 inhibitor developed by GSK [74]. It is approved only for treatment of adult patients with metastatic melanoma, with specific mutations of the BRAF gene, which encodes a Ser/Thr kinase B-Raf.

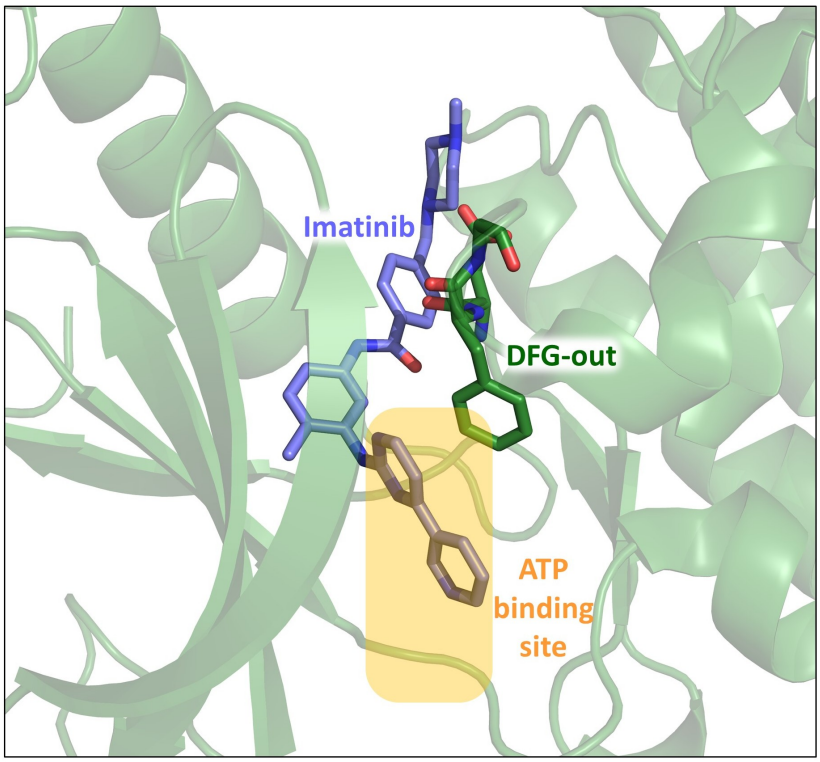


FIGURE 3.3: DDR1 kinase inhibitor imatinib binds at the active site however extends into another pocket which is accessible in the DFG-out conformation. Structure from PDB ID 4BKJ.

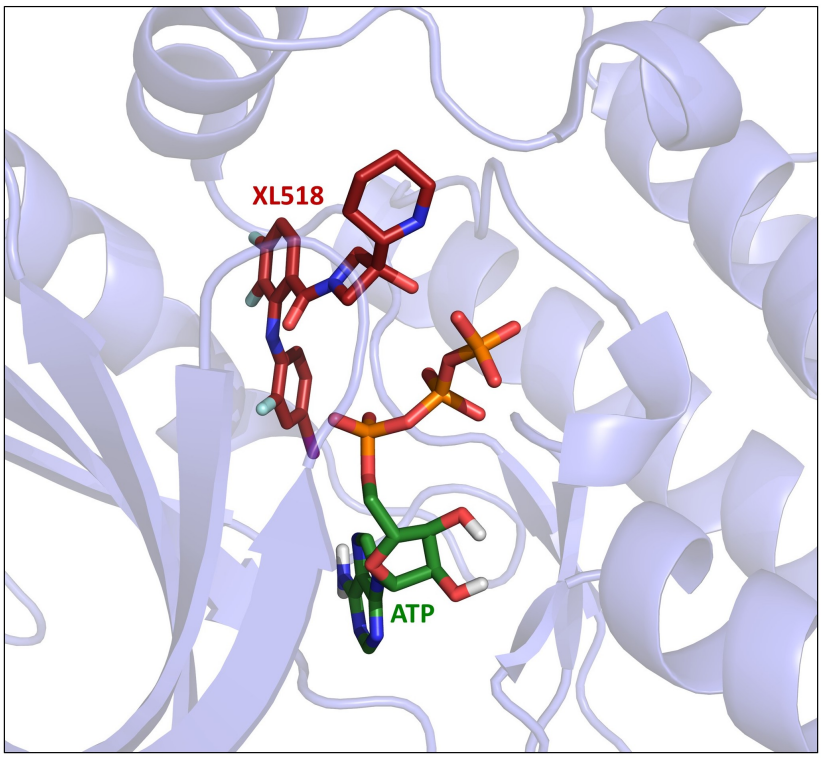


FIGURE 3.4: MEK1 kinase inhibitor XL518 (PDB ID 4AN2) bound at an allosteric site directly adjacent to the ATP binding site.

Further allosteric inhibitors are termed as class IV, which bind to a site completely separated from the ATP site. Various sites have been discovered, on both the N and C terminal lobes, depending on the kinase.

These distant allosteric sites are now known for many kinases, and both small molecule and peptide inhibitors and activators have been discovered [75]. Figure 3.5 highlights some of these known sites, overlaid on a structure of an AGC kinase, phosphoinositide-dependent kinase-1 (PDK1). The allosteric site for PDK1 is shown later, in figure 3.11. In this figure, ATP is bound to the active site in between the N and C terminal lobes, along with two Mg^{2+} ions. In red, a peptide inhibitor for the epidermal growth factor receptor (EGFR) kinase domain [76]. In green, compound **38** showed inhibition of checkpoint kinase Chk1, and binds to a site located around 13 Å from the active site, and with an IC_{50} of 1.3 μM [75]. A different allosteric site has been discovered for Abl kinase, highlighted in purple, [77].

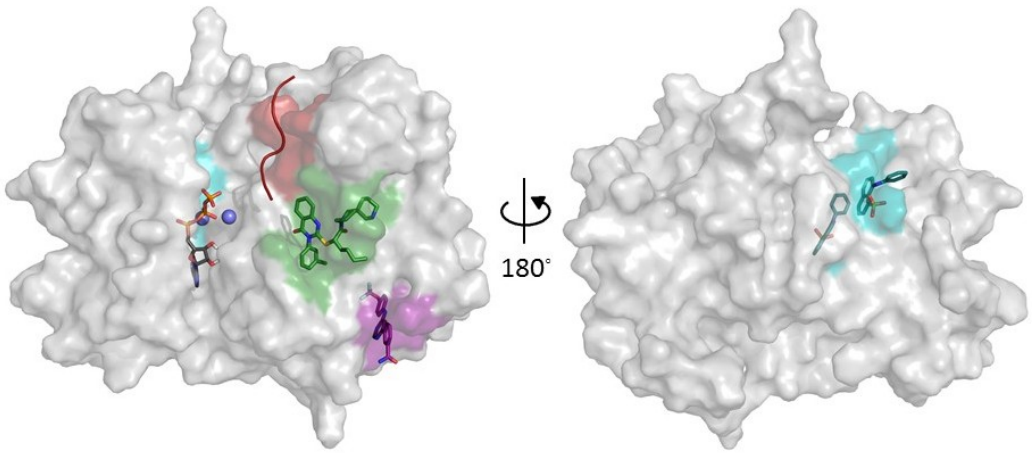


FIGURE 3.5: Structure of PDK1 with ATP and two Mg^{2+} bound at active site, and known allosteric sites for other kinases overlaid. Red: PDB ID 4R3R. Green: PDB ID 3F9N. Purple: 3K5V. Teal: 3PXZ.

In teal in figure 3.5, an allosteric site in the N terminal lobe was discovered for Cyclin-dependent kinase 2 (CDK2) [78], however in this case two molecules bind to this region, and require a large shift in the position of the α-helix C to accommodate one of them, as shown in figure 3.6. In this figure, the teal coloured helix C represents the shifted helix, with two molecules of

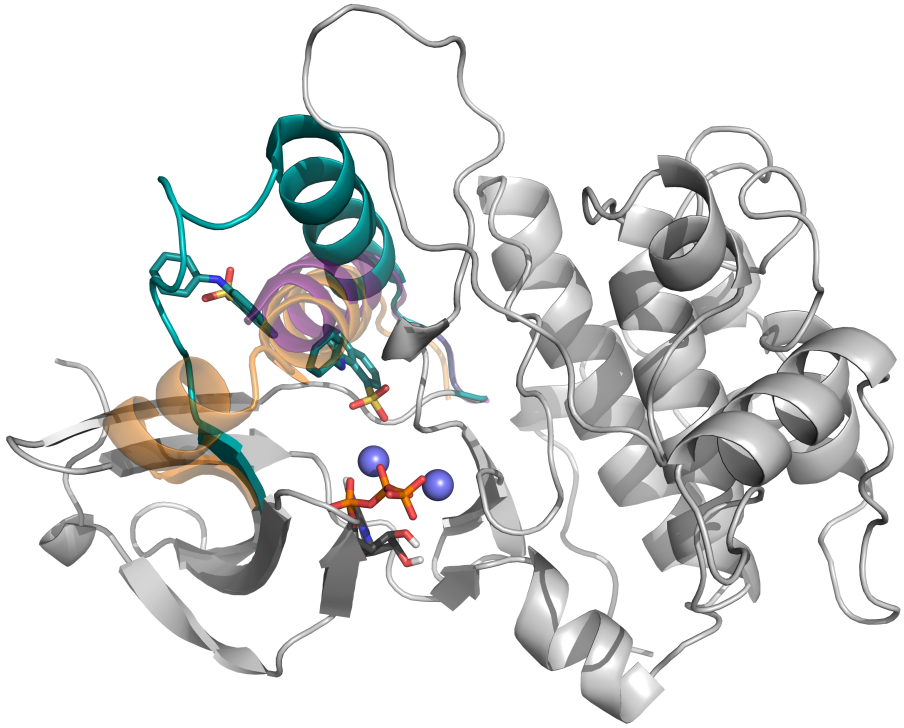


FIGURE 3.6: CDK2 allosteric site highlighting shift of helix C (See Figure 3.1 for kinase structural features). Teal: crystal structure PDB ID 3PXZ with two allosteric ligands. Purple: CDK2 without allosteric ligand (PDB ID 3MY5). Orange: Helix C in PDK1. Grey: Both overlaid on structure of PDK1.

inhibitor bound. In orange, the α-helix C from PDK1 is shown, and in purple the helix position in another structure of CDK2 without allosteric ligand (PDB ID 3MY5).

As of yet, there are no type IV kinase inhibitor on the market, however they do offer a promising solution to the problems faced by the other three classes of kinase inhibitor. As mentioned in section 1.3, evolution of enzymes relies on the conservation of the active site. Residues which are directly involved in the enzyme catalysed reaction are less likely to tolerate mutations, so often the active site and surrounding region are highly conserved. However if we target sites located away from this conserved region, it is potentially easier to find mutations which would allow selective binding, as there is a higher chance that some residues vary between family members. Also, as with type III inhibitors, there is the possibility to tune

the inhibition. As these sites are non-competitive with ATP, it is possible to achieve a range of inhibition depending on the conformation that the drug can stabilise. It is even possible to *increase* the activity, as this can also result in a desired therapeutic effect in some kinases [79]. In the case of active site compounds for kinases or any other enzyme, direct activation is not possible, as the endogenous ligand binding is blocked.

Particularly an issue with kinase inhibition is that concentrations of ATP in the cell are reasonably high (1 to 5 mM [80]). Therefore any active site compound has to both succeed at competing with high ATP concentrations, *and* do this selectively.

The IC_{50} of a drug is the concentration of drug required to obtain 50% inhibition. However as active site kinase inhibitors are also competing with ATP, the IC_{50} not only a measure of the drug affinity, but also on the concentrations of ATP.

$$IC_{50} = K_i(\frac{1 + [ATP]}{K_{M,ATP}}) \tag{3.1}$$

Where K_i is the inhibition constant of the ligand, [ATP] is the concentration of ATP, and $K_{M,ATP}$ is the Michaelis constant for ATP. For the high cellular concentrations of ATP this then contributes to the difficulty of developing good active site inhibitors.

Since type IV allosteric sites do not compete with ATP, this is not an issue. However a challenge with these type IV allosteric sites, is that as most of these sites are effectively protein-protein interfaces (PPIs), they tend to be far more shallow than the active site, leading to difficulties with potency due to insufficient ligand affinity. An interesting approach to solve this solution is the development of covalent allosteric drugs [81].

Covalent drugs are those which contain reactive groups which can bond irreversibly with their target proteins. They have been in use for some time, and in fact many were discovered even before their mode of action was understood [82], however toxicity concerns have resulted in reduced interest the development of new compounds [83]. With allosteric sites, toxicity could be less of an issue, provided that off target covalent binding can be avoided. This is easier with allosteric sites if non-conserved residues are the

target of the covalent modification, as this then reduces the chances of binding to a related protein [81]. As drugs bind covalently, they do not require the strong affinities of non-covalent drugs which rely only on non-bonded interactions to maintain a stable drug-protein complex.

The selectivity improvement seen by targeting allosteric sites can also be a source of a potential problem. As allosteric sites do not show the same evolutionary pressure to remain conserved, they are more prone to tolerating mutations, and this is a benefit in terms of selectivity. However this also could lead to subsequent issues with resistance, as if residues within the allosteric site mutate, the drug may no longer be effective [84, 85]. Yet resistance is also an issue for active site compounds, and so should not result in reduced focus on targeting allosteric sites [86].

3.1.2 PDK1 3-phosphoinositide-dependent protein kinase-1

PDK1 (or PDPK1) is a master AGC kinase. "Master kinase" in that it phosphorylates other kinases (at serine or threonine residues), of which it is part of the same family. In most kinases, a hydrophobic pocket exists within the N-terminal lobe (highlighted in purple in Figure 3.7), and this can interact with the hydrophobic motif (HM) which extends from the C-terminal lobe of the same molecule. With PDK1, this hydrophobic motif is not present. Instead, the hydrophobic pocket of PDK1 is involved when binding the substrate protein, where the hydrophobic motif belonging to the substrate can interact with the hydrophobic pocket of PDK1, and this instigates the allosteric event, as shown in figure 3.7. The terminal region of the substrate HM has been termed the "PIF" region, or "PDK1 Interacting Fragment". Therefore the allosteric site which this binds to in PDK1 is named the "PIF-binding pocket".

Substrate proteins must already be phosphorylated at their HM to increase recruitment by PDK1 [65, 66]. This phosphate interacts with a pocket directly adjacent to the PIF pocket [87], which facilitates the PIF binding to PDK1. This binds to the catalytically active conformation of PDK1 and so allosterically activates PDK1, where the active site can then catalyse the transfer of a phosphate to the substrate activation loop. Studies have also found

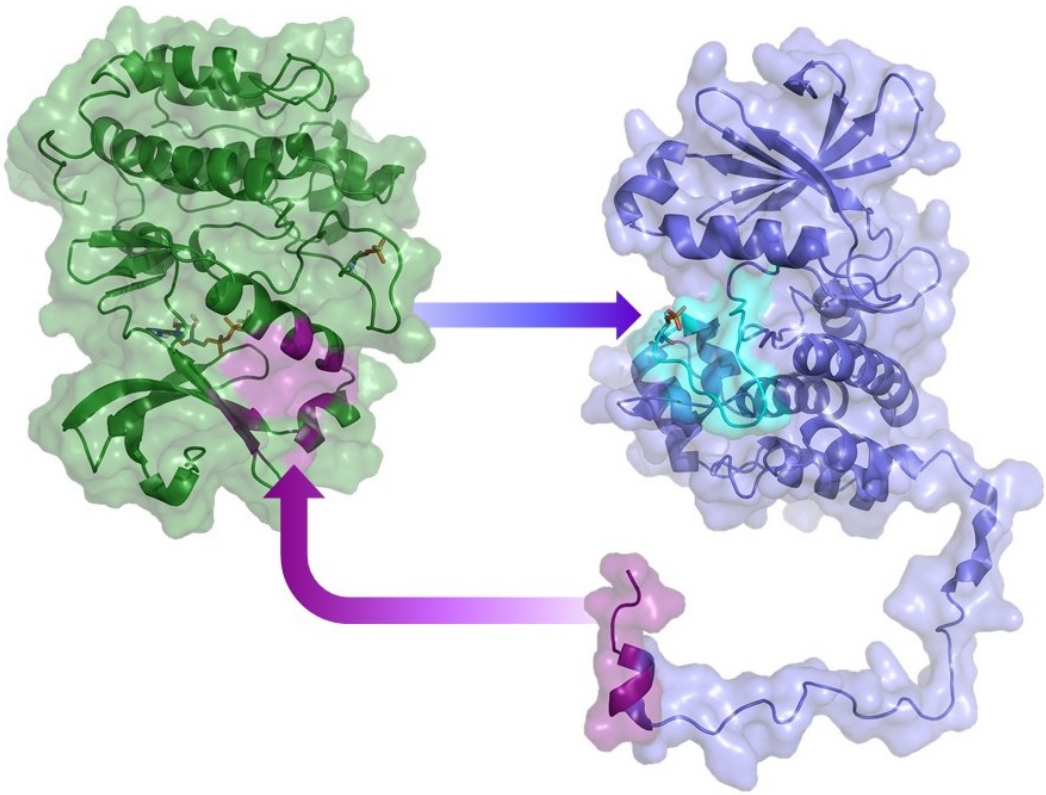


FIGURE 3.7: PDK1 (green), highlighting the PIF-pocket (purple). ATP bound to active site, and P-Ser169 shown in sticks. Substrate protein Akt in blue (PDB ID 1O6L altered to illustrate extended HM region). PDK1 interacting fragment (PIF) of Akt higlighted in purple. Activation loop of Akt in cyan, with P-Thr which is phosphorylated by PDK1 shown in sticks.

the inverse allosteric effect can occur; in that binding of a small molecule inhibitor at the active site can affect the binding a the PIF-peptide at the allosteric site [88].

The structure of the PDK1 catalytic domain is illustrated in figure 3.1. In addition to the catalytic domain, PDK1 possesses a pleckstrin homology (PH) domain, which can recruit PDK1 to the membrane during specific signalling events [89]. PDK1 is inherently active, and [89] autophosphorylates at Ser169.

Initial discoveries of allosteric modulators of PDK1 were peptides [90]. Inhibitors were only discovered as small molecules. This may be due to the far more specific interactions of a small molecule, whereas the sum of the activating and inhibiting interactions of the peptides do not result in inhibition as suggested by Sadowsky et al [91], who stated that:

> *"It is possible that the activation of PDK1 observed by PIFtides represents the net sum of interactions between a number of subsites, some of which can be activating and some inhibiting. Perhaps the small fragments can occupy individual subsites and thereby generate more potent activat- ing or inhibiting effects on PDK1."*

The structures used as a basis for all simulations of PDK1 were developed by Sadowsky et al. [91], and include various covalently bound small molecules (Figure 3.16) which are attached to the PIF pocket of PDK1. This was achieved by an artificial modification of a residue within this pocket to a cysteine, and a ligand could then be attached by a disulphide bond, in a process known as "disulphide trapping". Various positions within the PIF pocket were mutated, and one particular system was chosen for study (T148C), as this was more selective in binding. To this cysteine residue, several small molecules have been attached via a disulphide bond. Both activator and inhibitor small molecules bound to PDK1 T148C were developed, and activity of the PDK1-disulphide complex was determined with a kinase activity assay, using a substrate derived from the activation loop of a

substrate protein, Akt. This substrate is likely to only bind at the active (orthosteric) site, as the sequence which was used in the assay (KTFAGTPEYLAPEVRR) would not be long enough to bind to both allosteric and orthosteric sites simultaneously. Phosphorylation occurs at the threonine residue which is fairly central in this peptide chain, and this residue must be in close proximity to ATP for phosphoryl transfer to occur. PDK1 binds ATP at the orthosteric site, along with two Mg^{2+} ions which are crucial for catalysis. In the crystal structures from Sadowsky et al., a non-reactive ATP competitive ligand is present in the orthosteric site instead of ATP. Further crystal structures were therefore considered in order to correctly place ATP and magnesium ions within the binding site. In this, and many other kinases, each magnesium ion tends to adopt octahedral coordination geometry [92–94], with coordination to ATP, protein residues, and water molecules. The residues which coordinate to magnesium vary between kinases, however with PDK1 includes one residue of the "DFG motif" [95].

3.2 Methods

3.2.1 Molecular modelling

3.2.1.1 Ligand

The first considerations when setting up this model involved the parameterisation of the ligand, as it was necessary to correctly describe a non-standard disulphide bond, by which the ligands are attached to the PIF pocket. Parameters had to be set up for the ligand but include the correct description of the ligand-cysteine bond, which is not a standard residue included in either the GAFF [96] (General Amber force field; used for ligand) or Amber [97] (used for protein) force fields. To generate a suitable set of parameters, the X-ray crystallographic structures PDB IDs 3OTU (PDK1 with allosteric activator JS30), 3ORZ (PDK1 with allosteric activator 2A2) and 3ORX (PDK1 with allosteric inhibitor 1F8) were used as a starting point, and in each case, the coordinates of the allosteric ligand along with the Sγ and Cδ atoms from

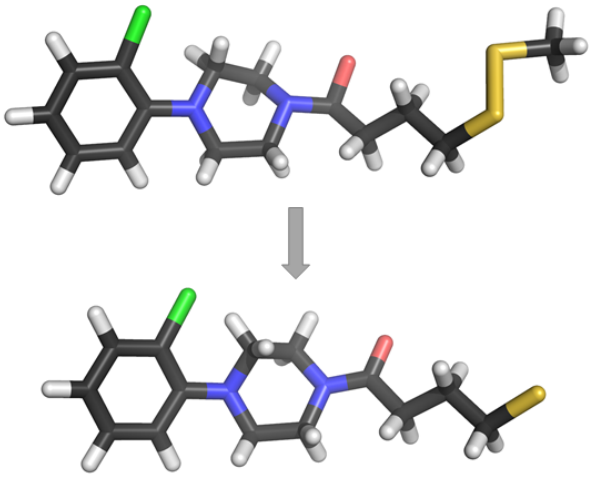


FIGURE 3.8: Ligand with atoms belonging to CYX residue, with Cβ replaced by a hydrogen atom, used to generate initial partial charges using Antechamber. Atoms belonging to CYX removed to balance charge to zero.

the Cys148 residue were retained, and Cβ was replaced with a hydrogen atom (Figure 3.8).

For the resulting thio-ether, partial charges could then be derived using the AM1-BCC methodology using Antechamber [98]. The partial charges initially assigned were modified to replicate the bond of a bound cysteine residue, as defined in the Amber force field. The atoms belonging to the Cys residue were then removed as in figure 3.8, and overall charge on the ligand was balanced (either to zero for uncharged, or +1 for protonated). The coordinates of the ligand were aligned to the X-ray crystal structure of the protein ligand complex using the software Pymol [99]. Three ligands were initially developed in this way: two agonists (JS30 and 2A2) and one antagonist (1F8).

Later this set was extended to the full set of compounds from the paper by Sadowsky et al. [91], which are based on the same scaffold as 2A2 and JS30, and are detailed in Tables 3.1 and 3.2. For these further compounds, no crystal structures were available, and so modifications to the ligand structures were made using 3ORZ and 3OTU as a template.

3.2.1.2 Protein preparation

Crystal structures were found in the PDB database, which were provided by Sadowsky et al. [91] for three systems. Two of these are of PDK1 bound to two different activating molecules: 3ORZ (ligand 2A2) and 3OTU (ligand JS30), and one bound to an inhibitor 3ORX (ligand 1F8). Experimental assays used a construct that contained residues 51–359 of the full wild type, so representing the catalytic domain of PDK1. Hence it would be ideal to model the same sequence which was used experimentally. In all cases, parts of the protein were missing; namely the activation loop (3ORX missing residue 237; 3ORZ missing residues 232-240; and 3OTU missing residues 232-236), and a section of the N-terminal region (3ORX missing residue 51-74; 3ORZ missing residues 51-73; and 3OTU missing residues 51-76). Model structures of PDK1 were set up for three systems (for ligands 2A2, JS30 and 1F8) using the software MODELLER [100]. The first models which were set up included the amino-acid sequence used in the kinase activity assay (residues 51-359 of the full wild type), meaning that part of the activation loop and part of the N-terminal region were modelled using the "automodel" function of MODELLER. Further models were constructed, again using the same method as above, but using a shorter amino acid sequence to remove this N-terminal region (removing residues 51-74). The model protein which was used for the apo-PDK1 was based on the crystal structure provided for the inhibited system, 3ORX. In all cases, crystal waters and ions were removed, and the protein structure was prepared using Maestro [101]: missing hydrogen atoms were added, and N-methyl and acetyl groups were added to the C and N terminal ends of the protein respectively.

3.2.1.3 Substrate peptide

The work from Sadowsky [91] provided activity data based on phosphorylation of a peptide substrate, which had been derived from the activation loop of a PDK1 substrate, Akt, with one mutation of a cysteine to alanine, in order to improve solubility. The full sequence of the peptide used

was KTFAGTPEYLAPEVRR, however for our simulation protocol this peptide was truncated, and modelled as an hexapeptide (KTFAGT) with an N-methyl amine group capping the C terminal end. This sequence seems sufficient, as when considering substrates of PDK1, these residues represent the most conserved part of the set of substrates (figure 3.9). Modelling the entire peptide would likely require longer simulation time, and increase the likeliness of incorrectly predicting the binding mode.

		PHOS								
Peptide	***K***	***T***	***F***	***A***	***G***	***T***	***P***	***E***	***Y***	***L***
Akt1	K	**T**	F	C	G	T	P	E	Y	L
Akt2	K	**T**	F	C	G	T	P	E	Y	L
Atk3	K	**T**	F	C	G	T	P	E	Y	L
RPS6KB1	H	**T**	F	C	G	T	I	E	Y	M
RPS6KA1	Y	**S**	F	C	G	T	V	E	Y	M
RPS6KA2	Y	**S**	F	C	G	T	I	E	Y	M
RPS6KA3	Y	**S**	F	C	G	T	V	E	Y	M
PRKACA	W	**T**	L	C	G	T	P	E	Y	L
PRKCZ	S	**T**	F	C	G	T	P	N	Y	I
SGK1	S	**T**	F	C	G	T	P	E	Y	L
SGK2	S	**T**	F	C	G	T	P	E	Y	L
SGK3	T	**T**	F	C	G	T	P	E	Y	L
PKN1	S	**T**	F	C	G	T	P	E	F	L
PKN2	S	**T**	F	C	G	T	P	E	F	L

FIGURE 3.9: Substrate peptide from Sadowsky paper, compared to activation loop of kinase substrates of PDK1. Colours represent residues which are most conserved across the set.

The peptide binding mode was predicted using the software Pepsite [102] as no crystal structure exists for the peptide used in the activity assay, or of PDK1 bound to a substrate kinase. Predicted structures were only reasonable when using the conformation of the protein from crystal structure with activator bound (3ORZ); and no reasonable predictions were obtained when using the inhibitor bound structure (3ORX). Reasonable predictions were determined by considering the distance of the γ-phosphate of ATP to the threonine of the peptide, as this is the residue which the phosphate from ATP would be transferred to.

3.2.1.4 ATP

The crystal structure which is being used to set up the model of PDK1 [91] has an ATP competitive ligand bound in the orthosteric site, rather than ATP. Therefore in order to obtain coordinates for ATP bound at this site, other crystal structures were used: PDB ID 4AW0 for the inhibited complex [103] , and PDB ID 4A07 for the activated complex [104] . The parameters and structure used for ATP were taken from work carried out by Meagher et al. [105].

3.2.1.5 Magnesium ions

Two magnesium ions are also located in the orthosteric site of PDK1, which are important in catalysing the transfer of a phosphate from ATP to the substrate protein. In the crystal structures provided by Sadowsky et al. [91], no magnesium ions are present in the binding site. Therefore coordinates for the two Mg^{2+} ions were also taken from crystal structures 4AW0 and 4A07. Several types of models for metal ions have been employed for protein simulations, which traditionally involved a spherical ion which is either bound [106] or unbound [107, 108] to the surrounding protein residues and ATP. More recent models are based on work by Aqvist [109] and attempt to better allow for changes in coordination during the simulation. To achieve this, is is possible to implement a "dummy model", where particles with partial charges (dummy atoms) are placed around the central metal ion, in the usual coordination for a particular metal ion. For the two Mg^{2+} ions in the orthosteric site of PDK1, parameters provided by Kamerlin et al. were used [110, 111] and converted for use with Amber, which describe an octahedral dummy model (Figure 3.10). A δ^+ charge is applied to each dummy atom, which must sum to the overall charge of the metal (2^+ for magnesium), and the central metal has $(n - 6\delta)$ charge (where n is the charge of the metal). Dummy atoms are bound to the metal centre, however no bonds are created between the dummy atoms and surrounding ligands, therefore changes in coordination during the simulation are possible. This model also shows improvement when two ions are in close proximity to one another, as with the

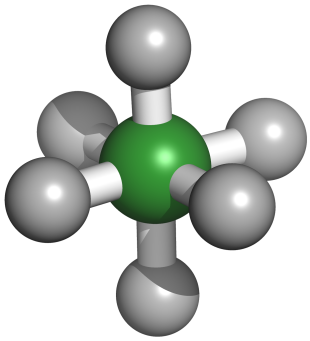


FIGURE 3.10: Octahedral dummy model for Mg^{2+} ion. Central green sphere is Mg ion and grey spheres are point charges arranged in octahedral geometry around central Mg.

non-bonded spherical model there can be issues with repulsion between the ions, which the dummy model should overcome. As the PDK1 system has two Mg^{2+} ions located close to one another within the ATP binding site, this model should show improvement over the previously used spherical model.

Placement of water molecules within the ATP binding site seems to be important, as distortions in both geometry of ATP and nearby residues occur if the Mg^{2+} ions are not sufficiently solvated prior to equilibration, to allow formation of an octahedral geometry. In other kinases, crystal structures with resolved water molecules in this site (PDB ID 1YTM (3 water molecules) and 1I59 (2 water molecules) show Mg^{2+} with octahedral geometry, as expected, with each Mg^{2+} interacting with oxygen phosphates of ATP, and with nearby residues and with either 2 or 3 water molecules. Two sets of simulations were initially set up, for PDK1 with inhibiting ligand 1F8: one which took crystal water molecules directly from PDB ID 4AW0, and another where water molecules were manually placed to allow each Mg^{2+} to achieve the octahedral geometry. Both allowed for the expected Mg^{2+} water coordination and so all further simulations were set up with water molecules manually placed at uncoordinated octahedral positions.

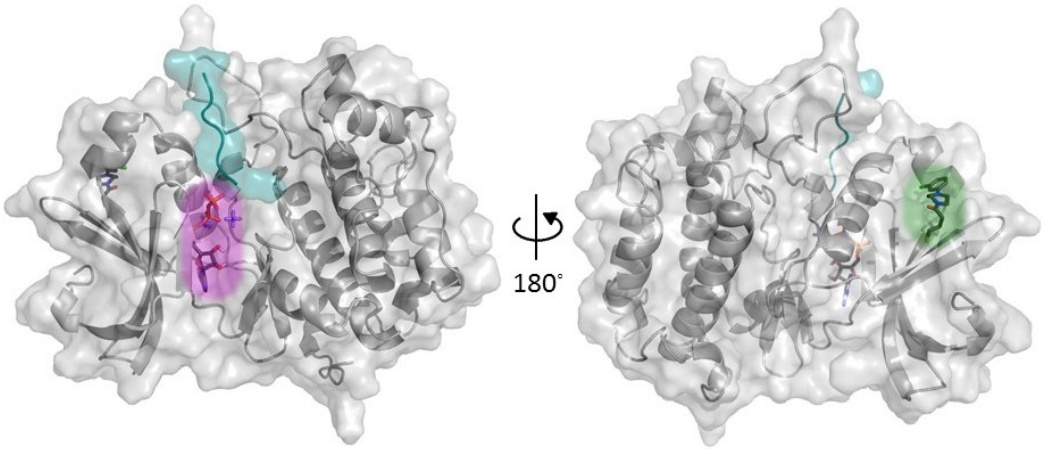


FIGURE 3.11: Structure of PDK1, highlighting ATP and dummy-model Mg^{2+} bound at the active site (purple), allosteric activator 2A2 bound at the PIF pocket (green), and substrate peptide (teal) predicted by Pepsite [102].

3.2.1.6 Ligand substrate protein complexes

In each case, a complex was set up as in figure 3.11 including protein, ligand, peptide, ATP and Mg^{2+} ions, using the software 'leap' from the Amber14 software package [98]. General Amber Force Field parameters were assigned to ligand atoms with the addition of the adapted disulphide bond parameters discussed previously, while the FF14SB-ILDN force field [45] was used to describe the protein. Phosphate parameters developed by Case et al. [112] were used to describe the phosphoserine located on the activation loop of PDK1. Magnesium parameters provided by Kamerlin et al. were used [110, 111]. Each model complex was then solvated in a box of TIP3P water molecules extending 10 Å from the edge of the solute, and Cl^- ions added to neutralise the net charge of the complex.

3.2.2 Molecular dynamics simulations

Three protein-ligand systems were initially considered and are referred to onwards as set A: PDK1 with activator molecule JS30, PDK1 with activator molecule 2A2, and PDK1 with inhibitor 1F8 (structures highlighted in

Figure 3.16 and Table 3.1). Also simulations of PDK1 without allosteric ligand bound were included in this set, allowing differences between bound and unbound proteins to be assessed. For the non-ligand bound simulation, the structure modelled from 3ORX was used for the protein. Later the set of simulations was extended to include further compounds, referred to as set B, and include compounds from the same scaffolds as JS30 and 2A2. Finally, simulations were run for the full compound set, and each set is shown in Table 3.2 in section 3.3.

The solvated models were energy minimised using sander, and equilibrated in NVT conditions using PMEMD (CUDA), from the software package Amber14 [98]. Energy minimisation using 200 steps of conjugate gradient with restraints on the solute was carried out in order to equilibrate only the solvent. This was then followed by 4500 steps of conjugate gradient with no restraints. Equilibration was carried out in the NVT ensemble, with harmonic Cartesian positional restraints of 10 kcal mol^{-1} $Å^{-1}$ initially on protein, ligand and ATP. Slow heating from 50 K to 300 K was done over 200 ps. Restraints on the protein and the ligand were retained while reducing those on ATP during the next 450 ps, and finally the restraints on protein and ligand were slowly reduced to zero (over 2.5 ns). A further 0.5 ns was run in the NPT ensemble in order to equilbrate solvent density. Equilibrium molecular dynamics simulations were run using the software SOMD (Sire/OpenMM) [56], for a simulation time of 1 μs using a 2 fs timestep and integration using the Leapfrog Verlet algorithm. Simulation was done in the NPT ensemble, using the Andersen thermostat [53]. Long range electrostatic interactions were calculated using the reaction field method [49, 113], with a cutoff of 10 Å and a reaction field dielectric constant of 78.3. Throughout all equilibration and production MD, SHAKE was applied to constrain all bonds involving hydrogen. In all simulations, snapshots were saved every 5 ps, resulting in 200k snapshots for every 1 μs simulation.

3.2.3 Computing distances

Individual distances were initially computed as distributions from each trajectory, to compare the set of simulations using KL or JS divergence. Distances were calculated using mdtraj [114], and a custom script was developed to output distributions of a selection of distances of interest. Distances were selected based on known structural features from the literature, and distance from the substrate peptide to ATP (see Figure 3.19).

3.2.4 Calculating dihedral angles

Torsional angles have been shown [115] to be a good way to describe correlated motions such as those involved in allostery, and could provide a more generalised approach to understanding allosteric motions, rather than motions in Cartesian coordinates. Also in the case where visible conformational changes do not occur, there may be subtle motions involving rotation around torsional angles, which are responsible for the activity, which may otherwise not be apparent. These analysis scripts calculate distributions of ψ, ϕ, $\chi1$ and $\chi2$, which are highlighted in figure 3.12.

3.2.4.1 KL divergence: dihedral angles

To compute KL divergence of torsional angles, a script was developed which utilises mdtraj [114], and calculates distributions for each ψ , ϕ, $\chi1$ and $\chi2$ angle. For each angle, a fixed value is then split over all empty histogram bins, and a file saved with the histogram data for each angle. This is then computed for each ligand bound simulation, and again for the simulation with no ligand bound. Another script then calculates the KL divergence between any two simulations, and gives a KL value per angle. This is summed into "backbone" ($\psi + \phi$) and "sidechain" ($\chi1 + \chi2$). Six identical PDB structures are then edited: four for the individual torsions (ψ , ϕ, $\chi1$ and $\chi2$); and then two for the summed backbone and sidechain. For each the KL value is input into the B-factor column of the PDB file. These are then loaded into Pymol [99] for visualisation. A flowchart of this process can be seen in figure 3.13.

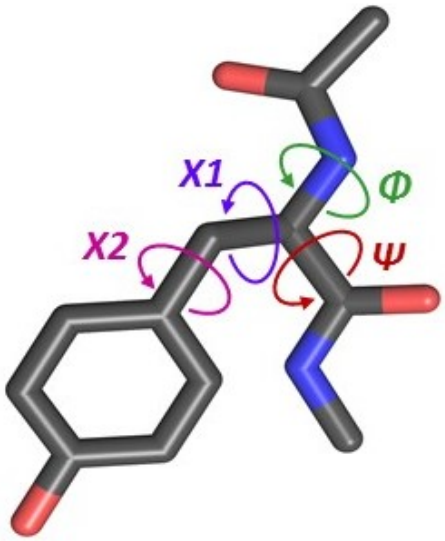


FIGURE 3.12: Torsional angles calculated illustrated on a tyrosine residue. Angles ϕ and ψ are backbone torsions, and $\chi 1$ and $\chi 2$ are the first and second sidechain torsions.

3.2.5 PCA

PCA was initially applied to a combination of four trajectories (3ORX, 3ORZ, 3OTU and APO) using the PyEMMA software [116], selecting only Cα coordinates as input dimensions. Component loadings were computed which gave a value of how much of the variance comes from a particular atom. These loadings were then visualised by inputting this value as a B-factor, and visualised using a colour scale using PyMOL to highlight motion involved in each mode[99]. Structures corresponding to the maximum and minimum values for PC1 and PC2 were output for each system. In addition, the value per snapshot, and the distribution of PC1 and PC2 were obtained for each input system.

PCA was later extended to include sets B and C, shown in 3.2. Distributions of PC1 were output for each system, and JS divergence computed for each compound pair. JS values were then clustered as described in 3.2.8.

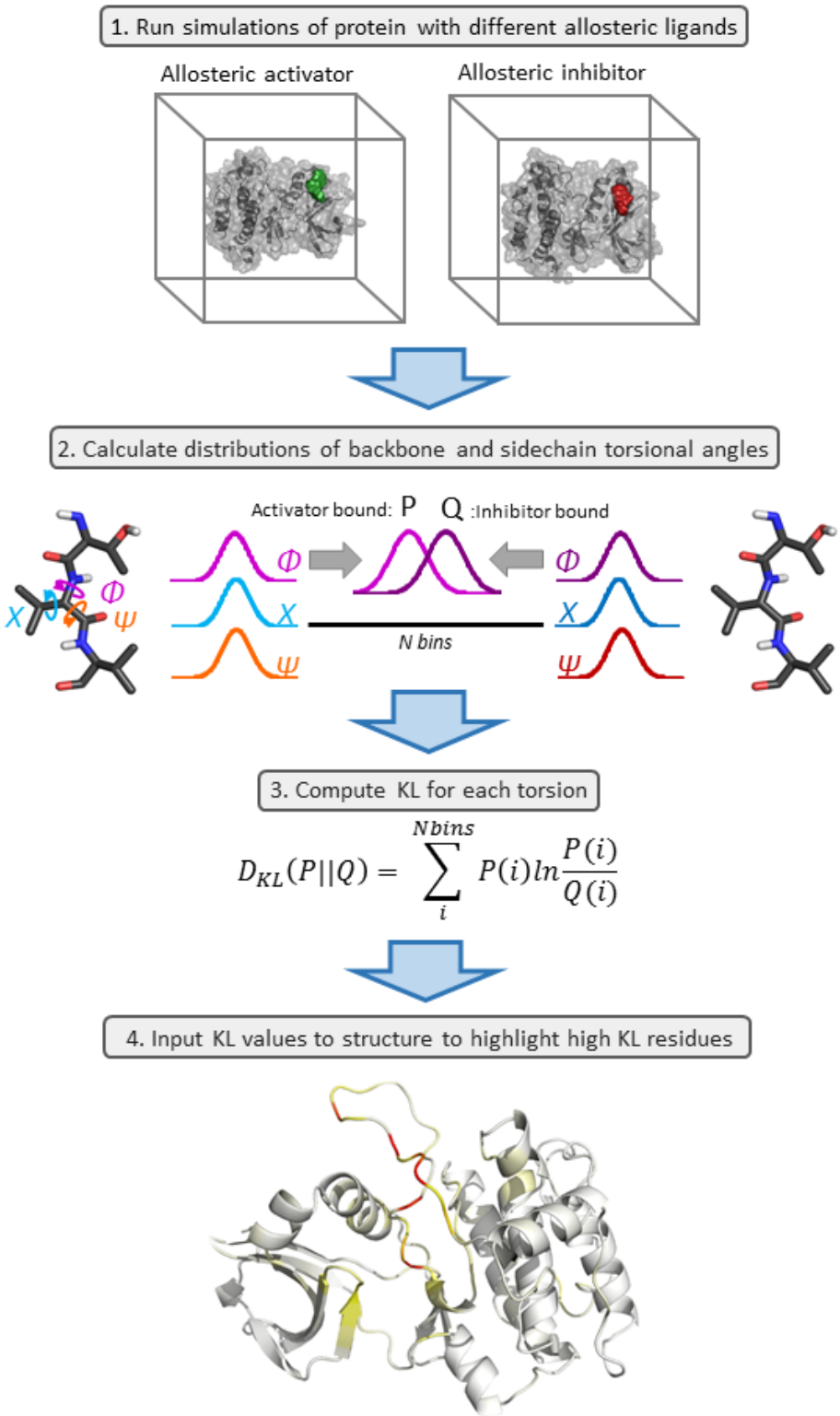


FIGURE 3.13: Calculation of dihedral KL.

3.2.6 Energy decomposition

Interaction energies were computed as described in section 2.4. Values are calculated for both the Lennard-Jones and Coulombic contributions at each timestep. These values can then be summed across a range of residues to give overall interaction energy profiles as probability distributions, for example for a ligand with the entire protein.

3.2.7 MI calculation

MI based on various descriptors was computed with three different python scripts: two were available or adapted from scikit [117, 118], and the third was an in house built script. For each descriptor previously calculated (PCA, distances, interaction energies) a file was saved with the value per snapshot for the descriptor, and this is given as input to the MI script. Further testing was completed with the third MI method ("in-house" script), using varying numbers of data points and number of bins, to determine optimal conditions. The process to calculate MI is shown in figure 3.14.

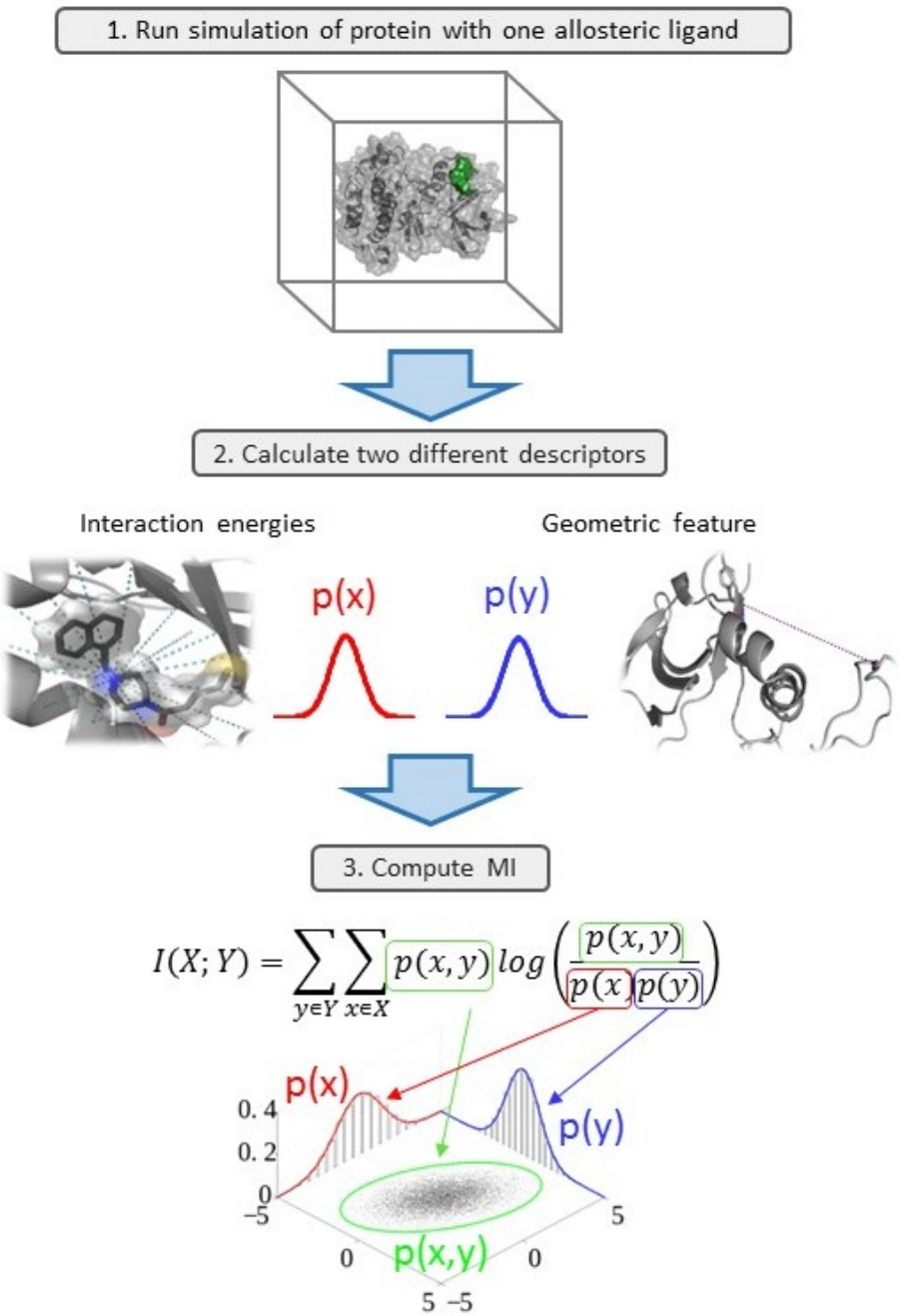


FIGURE 3.14: Workflow to compute MI between two descriptors.

3.2.8 Clustering of distance measurements and JS divergence

In order to easily identify similarities/differences between the JS divergence calculated for various metrics, spectral clustering [119, 120] was performed using the pyEMMA software [116]. A JS divergence matrix for each simulation pair was constructed, and used to devise a Gaussian diffusion kernel. The following equation defines the diffusion probability between two JS divergence values:

$$K_{ij} = e^{(-\frac{M_{ij}^2}{2\epsilon})} \tag{3.2}$$

In this equation, ϵ is a cutoff selected based on the structure of the eigenvalues, M_{ij} are the elements of the JS divergence matrix, and K is a measure of "distances" between two elements i and j. The eigenvalues and eigenvectors of a normalised matrix of K define how many clusters the dataset contains. For example, in figure 3.15, there is a separation in values between the first and second eigenvector, and so two clusters are selected.

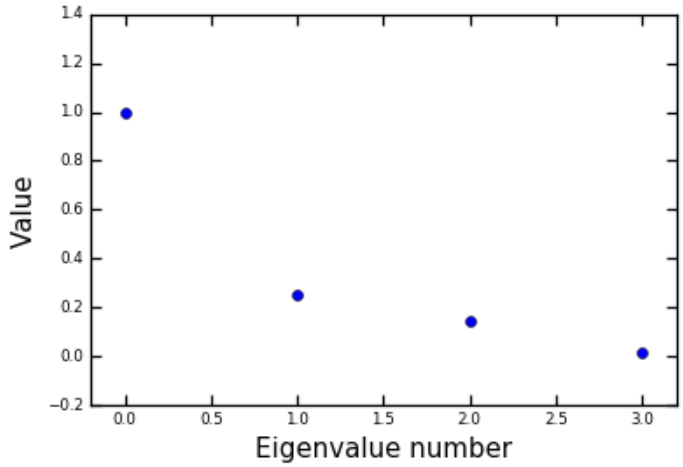


FIGURE 3.15: Eigenvalues for JS divergence values for a particular distance measured for four different simulations. ϵ is selected to maximise the difference in eigenvalues on trial and error basis. Two states are selected for this example as the separation in eigenvalues occurs between the first and second eigenvalue.

The PCCA algorithm was then applied to assign each simulation to a particular cluster, and results visualised using a matrix of JS values presented as a colour scale.

3.2.9 Availability of analysis scripts

Scripts to reproduce the results of this analysis, along with a tutorial using subsection of the full trajectory used in this thesis, are described in appendix B, and can be accessed on GitHub [121].

3.3 Results

Initially, simulations relating to the ligands for which crystal structures were provided were run and analysed, along with a simulation with no allosteric ligand. This was later extended to include further compounds in the paper provided by Sadowsky et al. [91], which were selected to give 4 compounds from each of scaffold A and B (figure 3.16) and also represent a range of activities. Later in the project, it was decided to extend this further to the entire set of compounds based on scaffolds A and B. Therefore, initial results are presented based on the 3 compound set, the extended 9 compound set, and then the full 24 compound set.

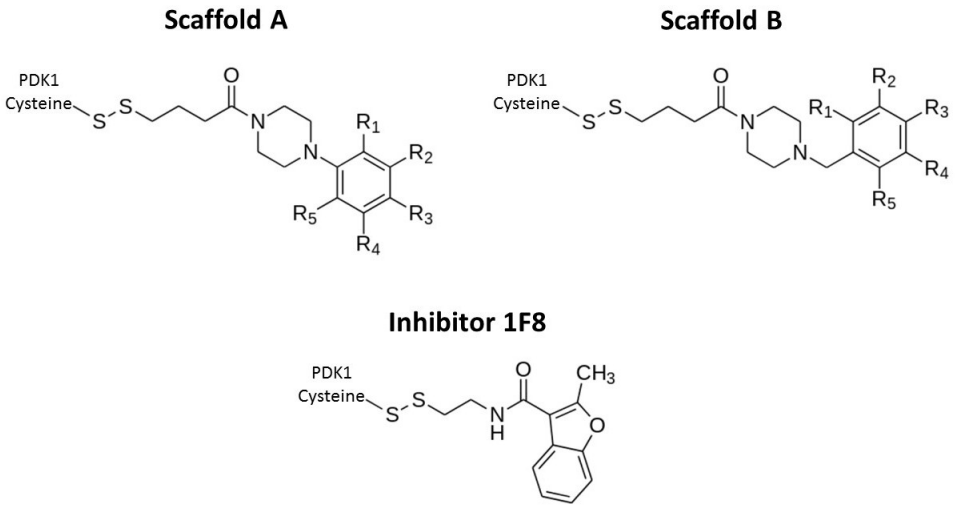


FIGURE 3.16: Scaffold A and B for allosteric activating ligands and structure of inhibitor 1F8. R groups shown in table 3.1.

Ligand	Scaffold	R1	R2	R3	R4	R5
JS12	A	H	H	OMe	H	H
JS26	A	H	OH	H	H	H
JS28	A	Cl	Cl	H	H	H
JS10	A	Cl	H	H	H	H
JS08	A	OMe	H	H	H	H
JS05	A	H	H	H	H	H
JS17	A	F	H	H	H	H
JS16	A	Me	Me	H	H	H
JS01	A	OH	H	H	H	H
JS25	A	H	OMe	H	H	H
JS15	A	H	H	F	H	H
JS14	A	Me	H	H	H	H
JS18	A	H	H	Cl	H	H
2A2*	A	H	Cl	H	H	H
JS19	A	H	H	CF_3	H	H
JS09	B	H	-OCH_2O-		H	H
JS04	B	Cl	H	H	H	Cl
JS02	B	H	F	H	H	H
JS03	B	Cl	H	Cl	H	H
JS21	B	H	Cl	Cl	H	H
JS23	B	H	Cl	H	H	H
JS24	B	H	H	Cl	H	H
JS30*	B	-CH=CH-CH=CH-		H	H	H

TABLE 3.1: R groups for scaffold A and B compounds. *Compounds 2A2 and JS30 correspond with crystal structures 3ORZ and 3OTU respectively.

3.3.1 Models

3.3.1.1 Protein

In all cases, the models obtained using the "automodel" function of the software MODELLER [100] showed a good fit to the crystal structures. In particular the missing loop residues in each case showed a reasonable conformation based on the adjacent residues which were present in the crystal structure. In the initial models (residues 51-359), the N-terminal region which

Set	**Scaffold**	**Compound**	**Activity**	**+/-**	Label
Set A	-	1F8 (3ORX)	32	2.2	1
	-	APO	100	-	2
	A	2A2 (3ORZ)	394	9.9	21
	B	JS30 (3OTU)	630	15	25
Set B	A	JS10	210	18	6
	A	JS18	370	32	18
	A	JS19	510	48	24
	B	JS09	240	22	9
	B	JS04	330	31	16
	B	JS23	460	39	22
Set C	A	JS12	160	21	3
	A	JS26	170	15	4
	A	JS28	200	20	5
	A	JS08	220	20	7
	A	JS05	220	18	8
	A	JS17	240	22	10
	A	JS16	240	22	11
	A	JS01	250	23	12
	A	JS25	260	23	13
	A	JS15	260	23	14
	A	JS14	270	24	15
	B	JS02	340	29	17
	B	JS03	380	32	19
	B	JS21	390	33	20
	B	JS24	460	73	23

TABLE 3.2: Full compound set with activities as percentage relative to apo. Compounds based on scaffolds A and B. Labels assigned as numbers in activity order, with compound **1** as inhibitor, and compound **25** as the most activating ligand JS30.

was missing from the PDB structure was modelled as an unstructured region (Figure 3.17 A).

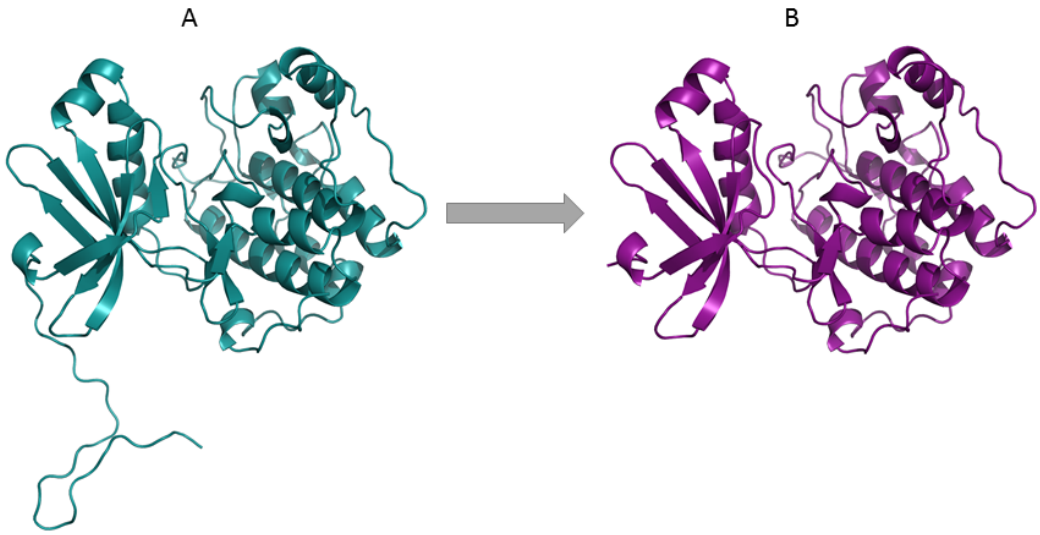


FIGURE 3.17: A) Initial model generated with sequence matching that used in the experimental assay by Sadowsky et al. [91] containing residues 51–359 of the full wild type. B) Shortened model used for all simulations using only residues 75-359 of the full wild type.

This extended out from the protein, and was deemed problematic for follow up MD simulations, as a much larger water box must be used to solvate. Furthermore, the significant flexibility of this N-terminal segment could pose sampling challenges. The main factor to consider when deciding the length of chain to model is whether this will have an effect on activity. This N-terminal region does not seem close enough in space to the active sites to interact. In addition, it is impossible to determine the accuracy of this part of the model, and no related crystal structures exist for this region. It is therefore sensible to omit this from further models given the difficulties this may cause with the simulation, when this modelled region may not even realistic, and so a shorter model consisting of residues 75-359 (with acetyl group added to residue 75 and N-methyl group added to residue 359), will be used to run simulations (Figure 3.17 B).

All further numbering of residues uses the first residue of the model as residue number 1.

3.3.1.2 Peptide

Peptide conformations predicted by Pepsite [102] seem to give reasonable results. As there are no structures of PDK1 with this peptide, or with substrate protein bound, related kinase structures are used to validate the results. Two related kinases were found which have available structures with active site peptide inhibitors. Protein kinase A (cAMP-dependent protein kinase) and protein kinase B (Akt) are both part of the same family of protein kianses as PDK1 (AGC kinases), and comparison of the predicted Pepsite conformation with crystal structures 1ATP and 3CQU suggests that the predicted conformation is reasonable as the predicted pose is close to these crystal strucutres, and allows the Thr of the peptide to be reasonably close to ATP, as seen in figure 3.18.

3.3.2 Distributions of distances relating to reaction mechanism

The initial analysis completed calculated distances between residues which are known in the literature to vary between activated and inhibited structures, and in addition the distance between ATP and the substrate. The distances computed are highlighted in figure 3.19.

3.3.2.1 ATP γ-phosphate to substrate Peptide-Thr distance

To determine whether particular structural changes influence the rate of phosphorylation, it is important to relate changes in the protein structure to some measurement which corresponds with substrate phosphorylation. Any of three distinct steps could be rate limiting during this process: substrate binding, phosphate transfer, or product release. Studies of PDK1 have shown that the phosphate transfer seems to be rate limiting [122], however, inhibition or activation by an allosteric ligand could, in theory, alter the rate of any of these three steps. Yet, for phosphorylation to occur, the substrate

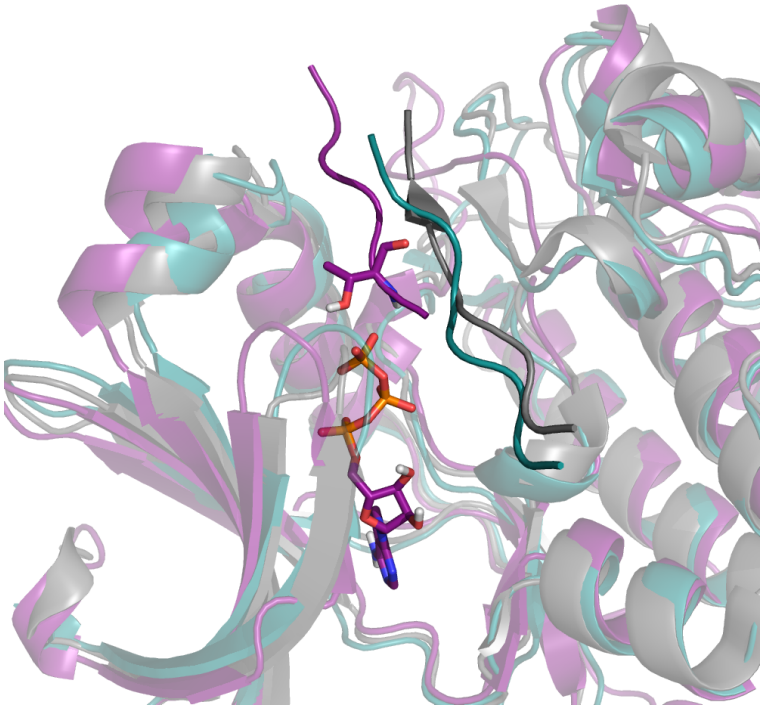


FIGURE 3.18: Predicted conformation of peptide using Pepsite [102] in purple. Crystal structure 3CQU in grey. Crystal structure 1ATP in teal.

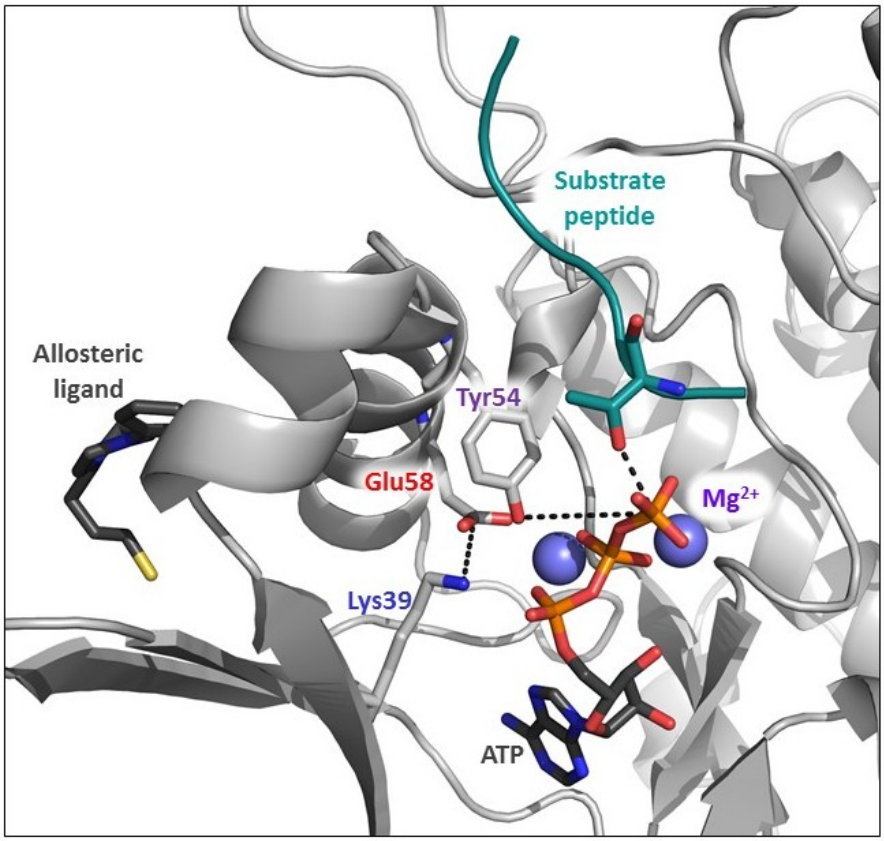


FIGURE 3.19: Distances computed based on information available in the literature. Distances calculated for Lys39(N) to Glu58(C); Tyr54(O) to ATP(γ-Phos); and substrate Thr(O) to ATP(γ-Phos).

System	Initial Distance (Å)	Average distance (Å)	Standard Deviation (σ)	KL divergence (apo as ref)
APO	3.33	3.83	0.43	0.00
1F8 (Inhibitor)	3.33	9.52	0.95	7.18
2A2 (Activator)	3.33	5.36	1.12	1.33
JS30 (Activator)	3.33	3.47	0.14	1.76

TABLE 3.3: Distances for peptide-Thr to γ-phosphate of ATP.

must bind such that the Thr residue which will be phosphorylated comes into reasonably close contact with the γ-phosphate of ATP, and as the phosphate transfer step is the rate limiting step, this could be monitored for different allosteric activator and inhibitor simulations. There are two potential mechanisms for phosphate transfer [123, 124], either via associative or dissociative pathways (figure 3.20). In both scenarios, the γ-phosphate of ATP must be within a reasonable distance (a few Å) to the substrate peptide threonine (Pep-Thr) for the reaction to proceed. Therefore it would be useful to see differences between activated and inhibited complexes, in the distance from the Pep-Thr, to the phosphate of ATP which will be transferred. Starting from the same peptide conformation with identical ATP to Pep-Thr distance, this was measured for the full 1 μs trajectory, and the results are shown in table 3.3 and figure 3.21. This shows that with the most activating compound (JS30), distances remained consistently at reasonably short distances, which would allow transfer of a phosphate. With activator 2A2, these distances were slightly longer, however reasonably long sections of the trajectory are at distances short enough for phosphate transfer to occur. For the inhibitor simulation, within around 200 ns, the peptide Thr is too far from the ATP site, and average distances are above 9 Å.

To validate these results, repeat runs of 100 ns were completed for the apo, inhibitor bound, and one activator bound (2A2) simulations. Results are in table 3.4 and highlight that in all cases, the values for the activator bound are shorter than inhibitor bound. It also highlights that both the activator and inhibitor bound have only small variations in the values between runs, with standard deviations of 0.16 Å and 0.67 Å respectively. However

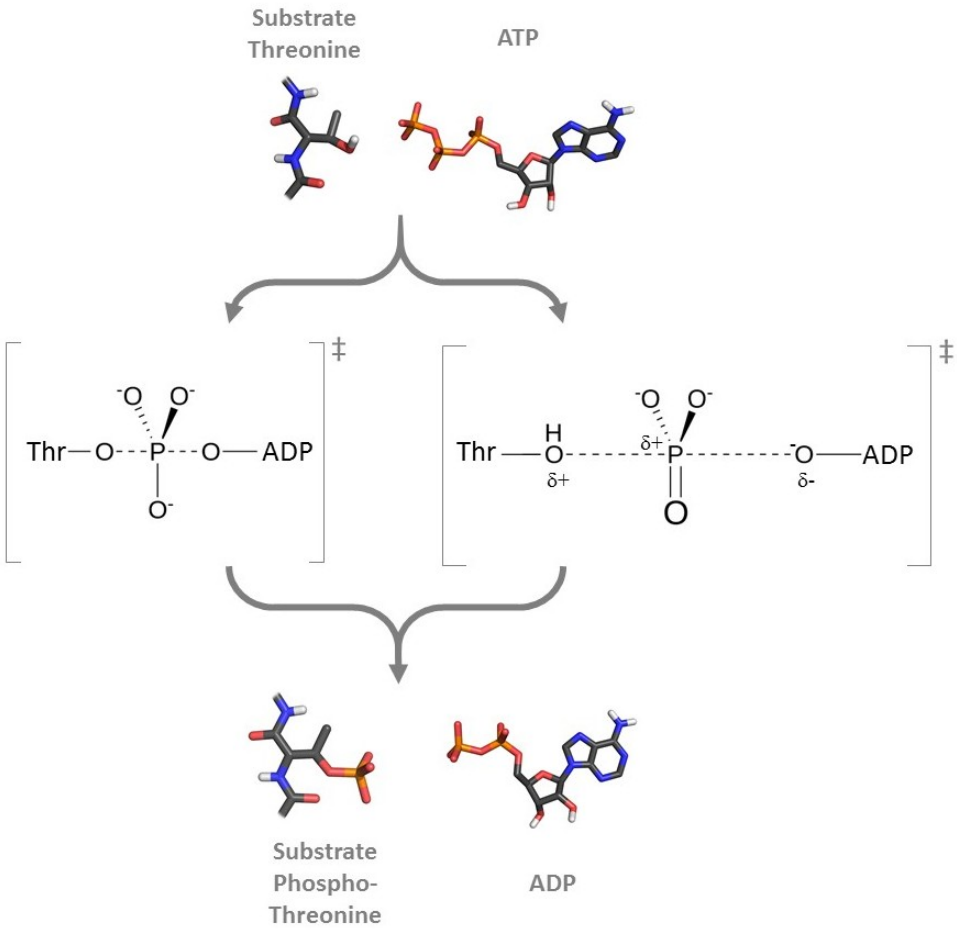


FIGURE 3.20: Phosphorylation of substrate kinase at serine, threonine, tyrosine, or histidine could occur via associative (S_N2-like: left) or dissociative (S_N1-like: right) mechanisms. Figure adapted from reference [124].

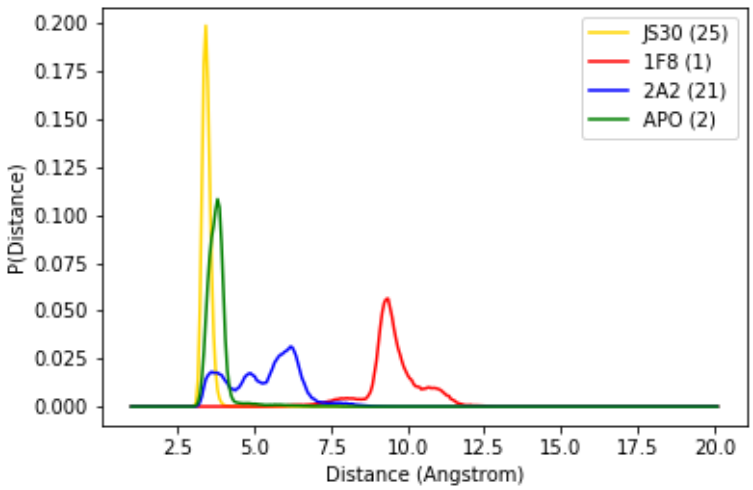


FIGURE 3.21: Distributions of substrate peptide Thr to γ-phosphate of ATP for the four original simulations completed. These are based on the 3 crystal structures provided for 3ORX (inhibitor 1F8), 3ORZ (activator 2A2) and 3OTU (activator JS30).

with the apo simulation, there is some variation, and standard deviation is much higher (2.39). This could be consistent with the trend in activity, as inhibitor bound is consistently at distances too long for catalysis to occur, activator bound stabilises the peptide close to the active site, and apo gives shorter distances *sometimes*, however not as consistently as activator bound.

System	Run 1	Run2	Run3	Run4	Run5	*Mean*	σ
APO	6.48	9.71	6.93	4.37	3.63	*6.22*	*2.39*
1F8 (Inhibitor)	8.20	6.83	7.74	7.13	8.39	*7.66*	*0.67*
2A2 (Activator)	3.58	3.37	3.51	3.55	3.81	*3.56*	*0.16*

TABLE 3.4: Average distances for repeat runs. Peptide-Thr to γ-phosphate of ATP.

The results for the substrate distance for the full compound set can be found in figure 3.22. In this plot, the orange horizontal line for each compound corresponds to the median, the box represents the lower and upper

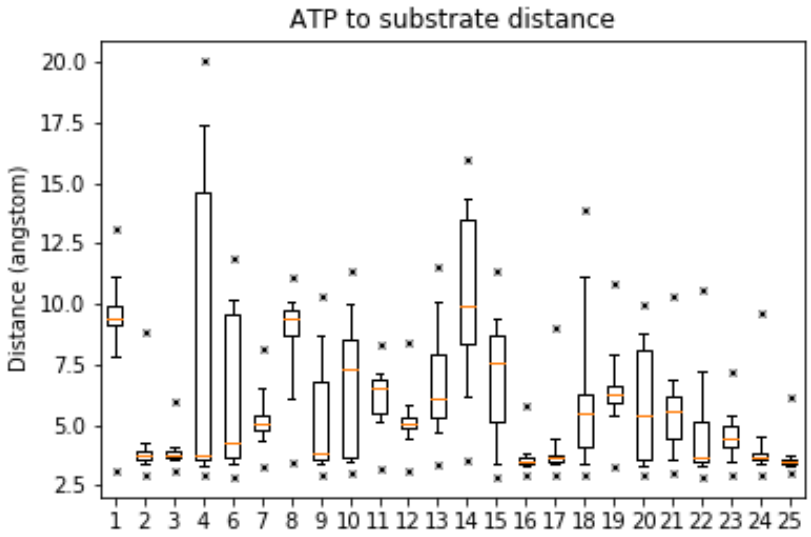


FIGURE 3.22: Distributions of the distance between the γ-phosphate of ATP and the Thr residue of the peptide which would be phosphorylated. All simulations are 1 μs. Compounds numbered as in table 3.2. JS14 is excluded as peptide dissociates from the active site in this simulation.

quartiles, and the "whiskers" (lines extending from each box) represent values at 5^{th} and 95^{th} percentile, and "x" marks the minimum and maximum values. From these distributions, the inhibitor bound simulation has the majority of distances above 7.5 Å. In most cases with activator bound, although some of the distribution lies at longer distances, there is significant probability at values short enough for phosphate transfer to occur. In the highest activating compounds (i.e. compounds 15-24 in figure 3.22), there is a much higher percentage of snapshots at shorter distances.

From figure 3.21, it is clear that the inhibitor bound simulation shows different behaviour, however this is more difficult to analyse when dealing with large sets of compounds. In order to more easily compare these distributions, the JS divergence was computed for each pair of compounds. In figure 3.23, the JS divergence for each of the four original simulations is shown. Darker colours represent higher JS divergence, and so larger variations between the distributions. White represents a JS divergence of zero,

and hence diagonal values are white, as $JSD(P \parallel P) = 0$.

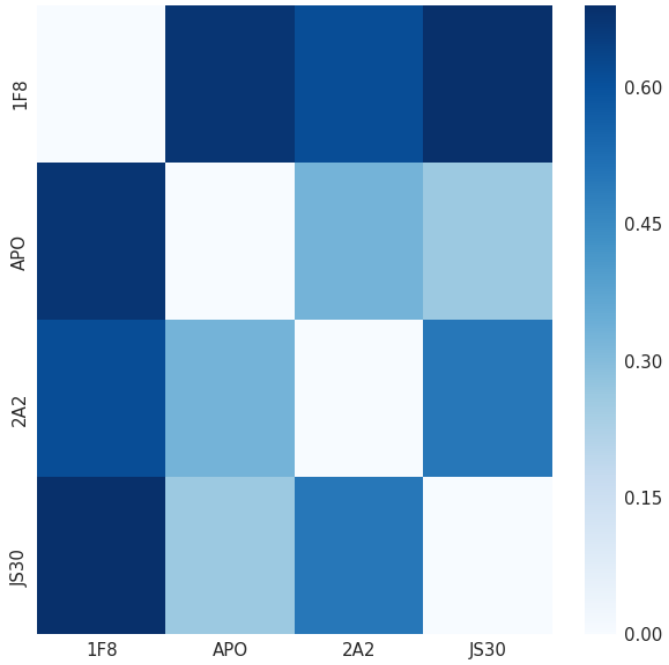


FIGURE 3.23: JS divergence for original four compound set of ATP-Peptide distance distributions. Clustering described in section 3.2.8, using ϵ=0.1 with 2 states.

This shows clearly that inhibitor 1F8 has the largest differences to any other compounds in this set, where JS divergence values when comparing 1F8 to every other simulation are around values of 0.6 or above.

This was then extended to the full set of compounds, and shown in figure 3.24. The results for this show some indication of a trend, with the highest activating compounds showing the most similarity.

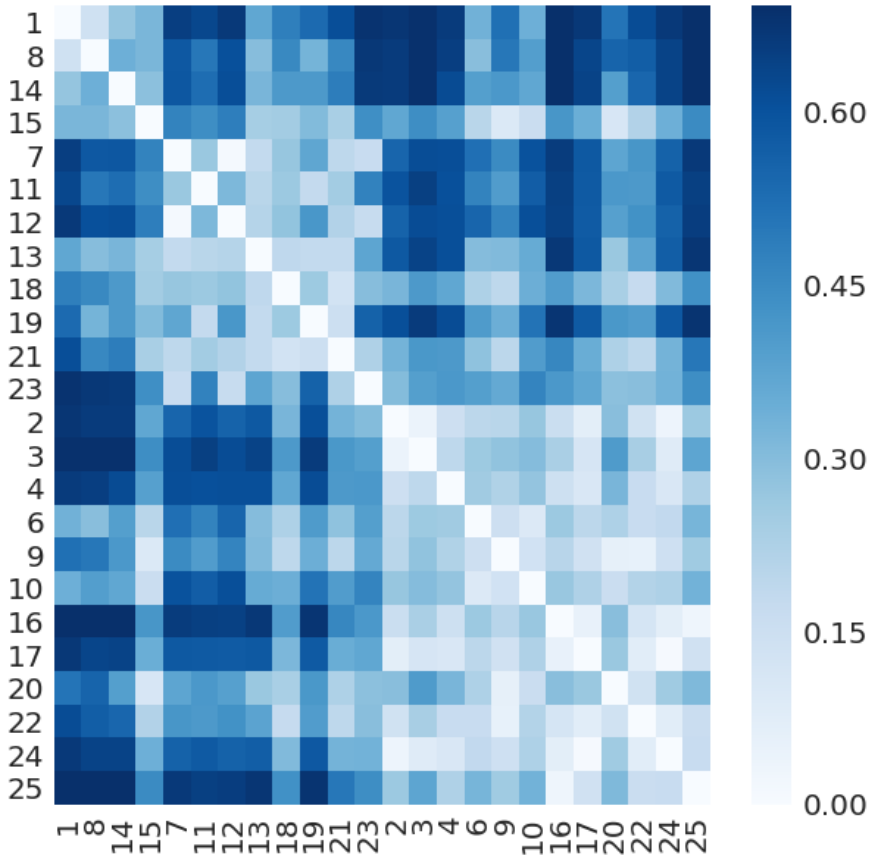


FIGURE 3.24: JS divergence for ATP-Peptide distance distributions. Clustering described in section 3.2.8, using ϵ=0.2 with 3 states.

3.3.2.2 Lys39 to Glu58 distance: a salt bridge between the active and allosteric sites.

In a study by Engel et al. [125], it was suggested that activation of PDK1 must involve a conformational change, which they state involves motion of the α-helix C, which is located between the PIF pocket and the ATP binding site (see Figure 3.1 and Figure 3.11). Movement of this helix results in motion of Glu58 (located on the helix), which in turn affects the position of another residue, Lys39, to interact with the phosphates of ATP. The distance between these two residues was measured by the distance between the sidechain N atom of lysine, and the terminal sidechain C atom of glutamic acid.

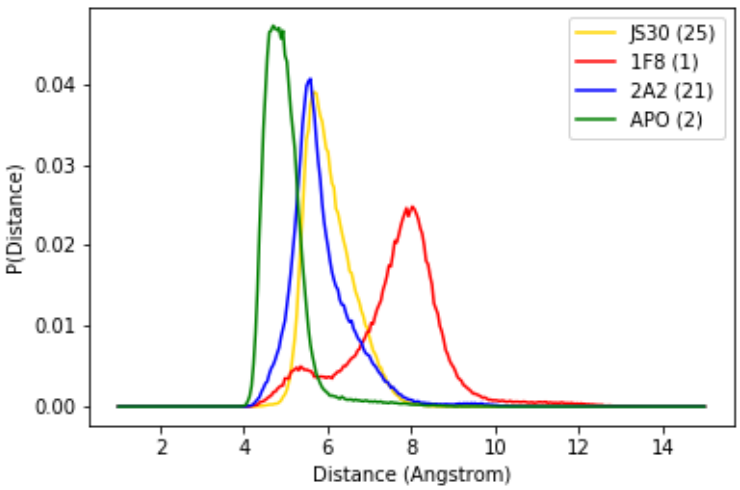


FIGURE 3.25: Distributions of the Lys39-Glu58 distance for the original set of simulations. Activator JS30 (3OTU), activator 2A2 (3ORZ), inhibitor 1F8 (3ORX) and with no allosteric ligand bound (APO).

A set of values for this distance was obtained over the course of the simulation, and the data plotted as a histogram (Figure 3.25). Average distances and standard deviations, along with KL divergence were computed and values are also summarised in table 3.5. Compounds 1F8 and JS30 show the largest differences in activity to Apo-PDK1, being the most inhibiting and

most activating ligands respectively. These two ligands also show the highest KL relative to Apo for this distance, which is consistent with the trend in activity. Compound 2A2 is also activating, but shows less change in activity relative to Apo and the KL value obtained is also smaller.

System	Initial Distance (Å)	Average distance (Å)	Standard Deviation (σ)	KL divergence (apo as ref)
APO	3.30	3.56	0.34	0.00
1F8 (Inhibitor)	3.62	5.26	0.77	2.58
2A2 (Activator)	4.05	4.13	0.49	1.30
JS30 (Activator)	3.86	4.25	0.36	3.31

TABLE 3.5: Distances from N of Lys39 to C of Glu58.

However when this was extended to the full set of compounds, the result is less clear. There is no clear trend related to activity, as shown in figure 3.26.

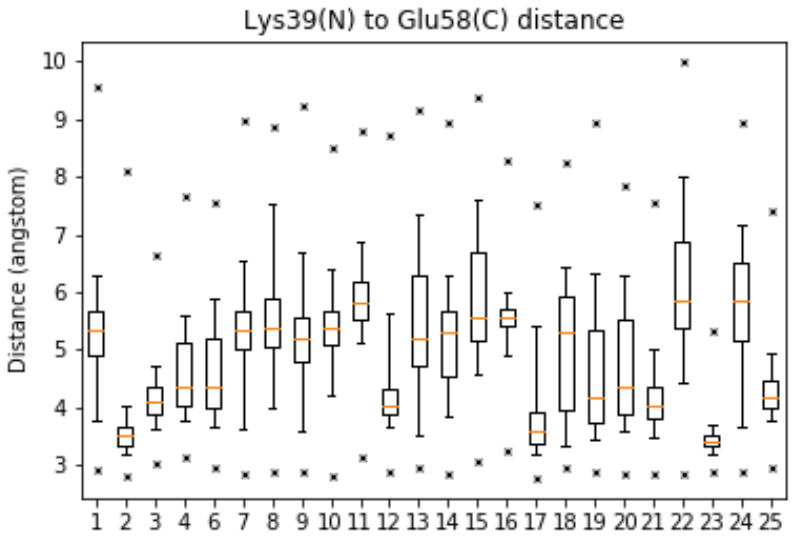


FIGURE 3.26: Distributions of the distance between Lys39 and Glu58. Compounds numbered as in table 3.2.

As with the previous distance computed, a matrix was first constructed for JS divergence values using the original set of four compounds, and can

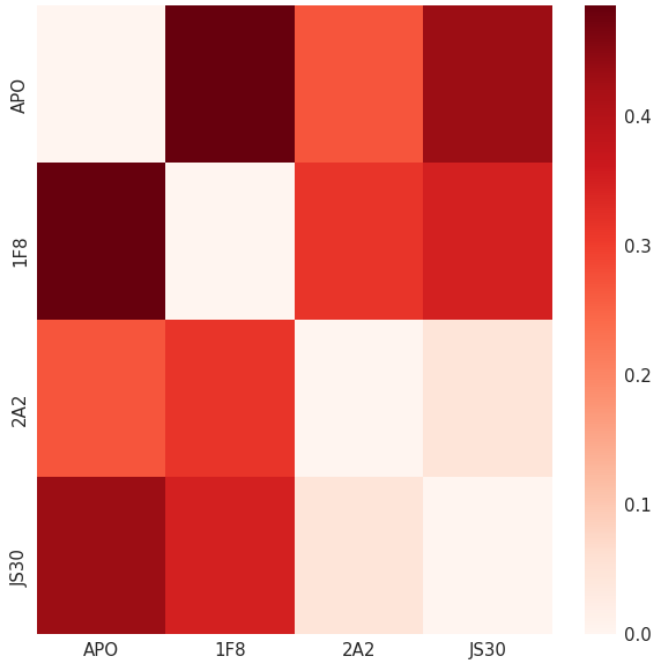


FIGURE 3.27: JS divergence for original four compound set of Glu-Lys distance distributions. Clustering described in section 3.2.8, using ϵ=0.2 with 2 states.

be seen in figure 3.27. This indicated similarity between the two activators, and showed that inhibitor 1F8 and apo were different to the two activators, but also different to each other. This was then extended to the full set, and it is clear that there seems to be no trend in these values based on activity for the full set of activators (figure 3.28), however some compounds show more similar behaviour to each other than to others.

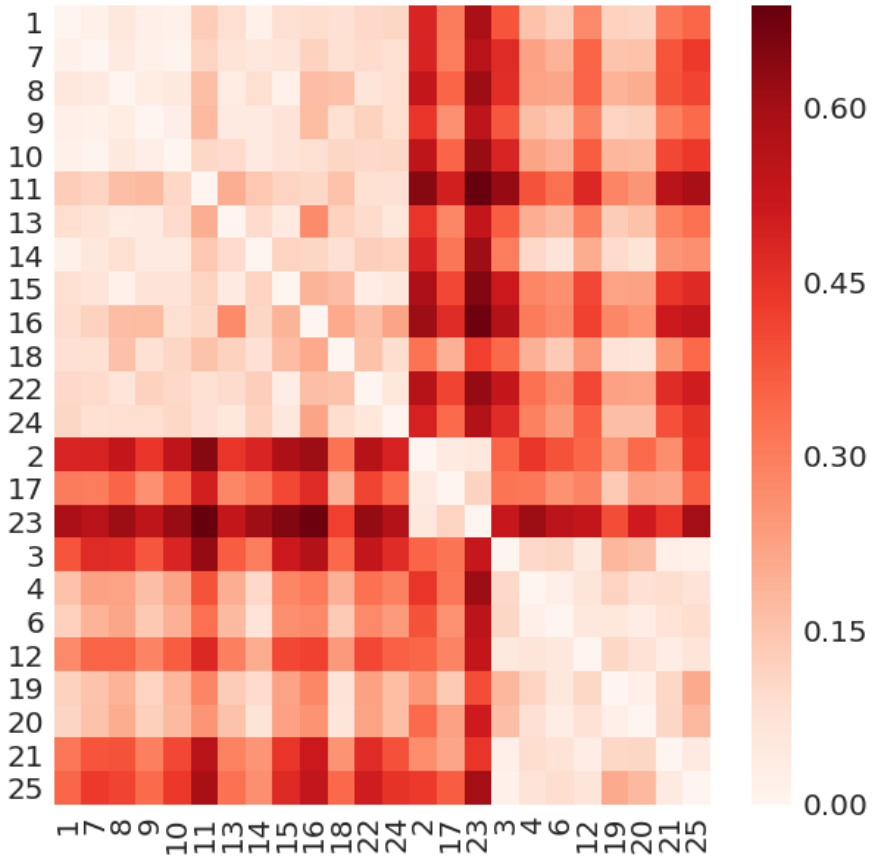


FIGURE 3.28: JS divergence for Glu-Lys distance distributions. Clustering described in section 3.2.8, using ϵ=0.1 with 4 states.

3.3.2.3 Tyr54 to ATP distance varies between activated and inhibited conformations.

Differences in the position of Tyr54 can be seen between activated and inhibited crystal structures (figure 3.29). In activated complexes, visual inspection of the X-ray structures suggests that the hydroxyl group of Tyr54 is close enough to ATP to form hydrogen bonds with γ-phosphate of ATP. In the inhibited complex, this side chain flips away and no H-bonding is possible.

This distance was monitored for 1 μs of simulation, and in the inhibited complex this distance remained in one distribution of values variation during the simulation, at relatively long distances, between 12 and 23 Å (Figure 3.30).

However with either of the two activators bound this group seems to be much more dynamic, and both have significant amount of the distribution as shorter distances (figure 3.30) than the inhibited complex. By comparison of the distribution of these values using KL divergence, it can be seen that relative to PDK1 with no allosteric effector, inhibitor bound PDK1 shows the largest KL value compared to the two activators (table 3.6). Differences in how residues interact with ATP could be key to explaining differences in activity, since stabilisation by hydrogen bonding of the phosphates of ATP could facilitate phosphate transfer. It is also important to highlight the shift to shorter values for the Apo simulation. As the starting structure for this simulation was the same as that for the inhibitor bound, the inhibitor does seem to stabilise the conformation of Tyr54 which results in longer Tyr54-ATP distances.

System	Initial Distance (Å)	Average distance (Å)	Standard Deviation (σ)	KL divergence (apo as ref)
APO	15.89	13.08	2.99	0.00
1F8 (Inhibitor)	15.89	16.81	1.23	5.80
2A2 (Activator)	5.58	13.70	5.28	1.32
JS30 (Activator)	5.80	12.73	2.64	1.43

TABLE 3.6: Distances for Tyr54(O) to γ-phosphate(P) of ATP.

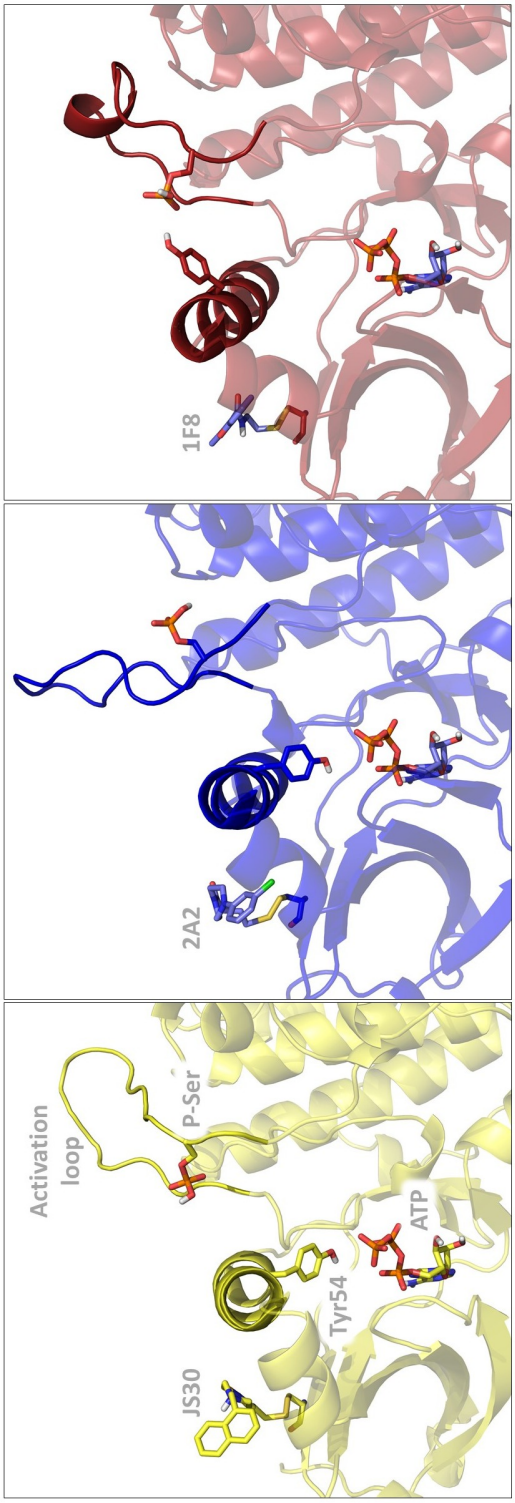


FIGURE 3.29: Conformations of Tyr54 in the PDB structures of 3OTU (activator JS30); 3ORZ (activator 2A2); and 3ORX (inhibitor 1F8).

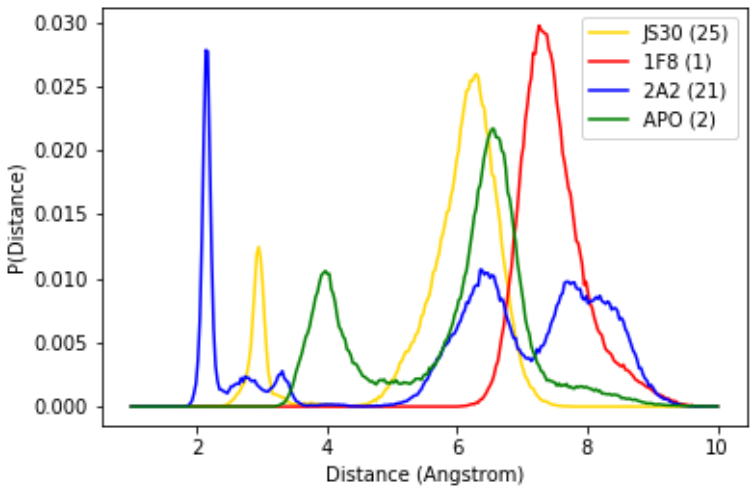


FIGURE 3.30: Distributions of Tyr54 (O(H)) to γ-phosphate (P) of ATP for the four original simulations completed. These are based on the 3 crystal structures provided for 3ORX (inhibitor 1F8), 3ORZ (activator 2A2) and 3OTU (activator JS30).

This was then extended to the full set, and results summarised in figure 3.31. This shows that only the inhibitor bound simulation has the majority of the distribution above distances of around 15 Å.

The JS divergence matrix on the original set of four compounds highlights the largest difference between the most activating, and most inhibiting ligands (figure 3.32).

Extended to the full set, some differences can be seen. The inhibitor bound (1F8) shows large differences with almost all of the set, while all activators show lower JS values with at least a subset of activators. There does not seem to be a trend relating to either the amount of activation or the scaffold. Some compounds show similar behaviour and have similar activities, such as JS26, JS10, JS08, JS17 and JS16, which all have activities in the range 170-240 % relative to apo.

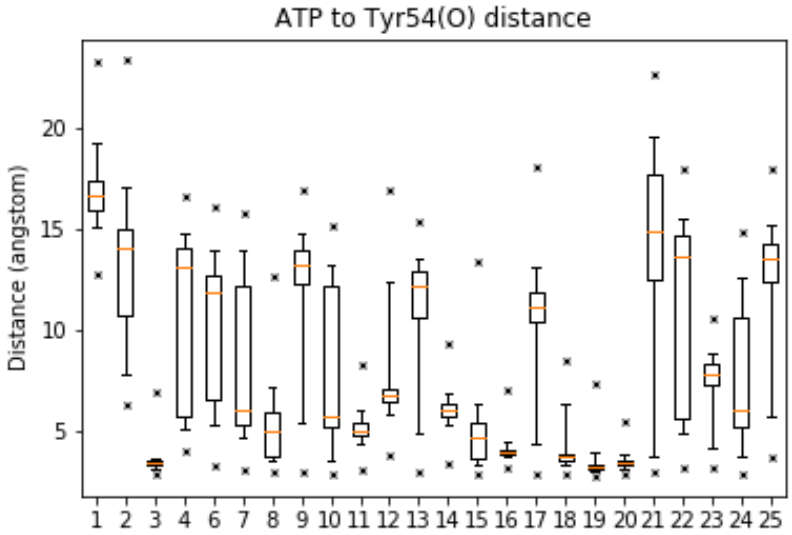


FIGURE 3.31: Distributions of the distance between the γ-phosphate of ATP and Tyr54. Compounds numbered as in table 3.2.

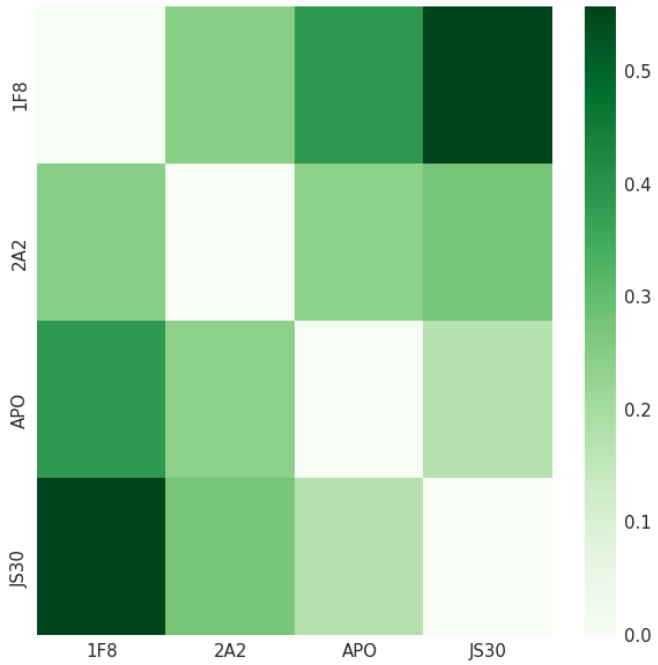


FIGURE 3.32: JS divergence for original four compound set of Tyr54-ATP distance distributions. Clustering described in section 3.2.8, using ϵ=0.2 with 2 states.

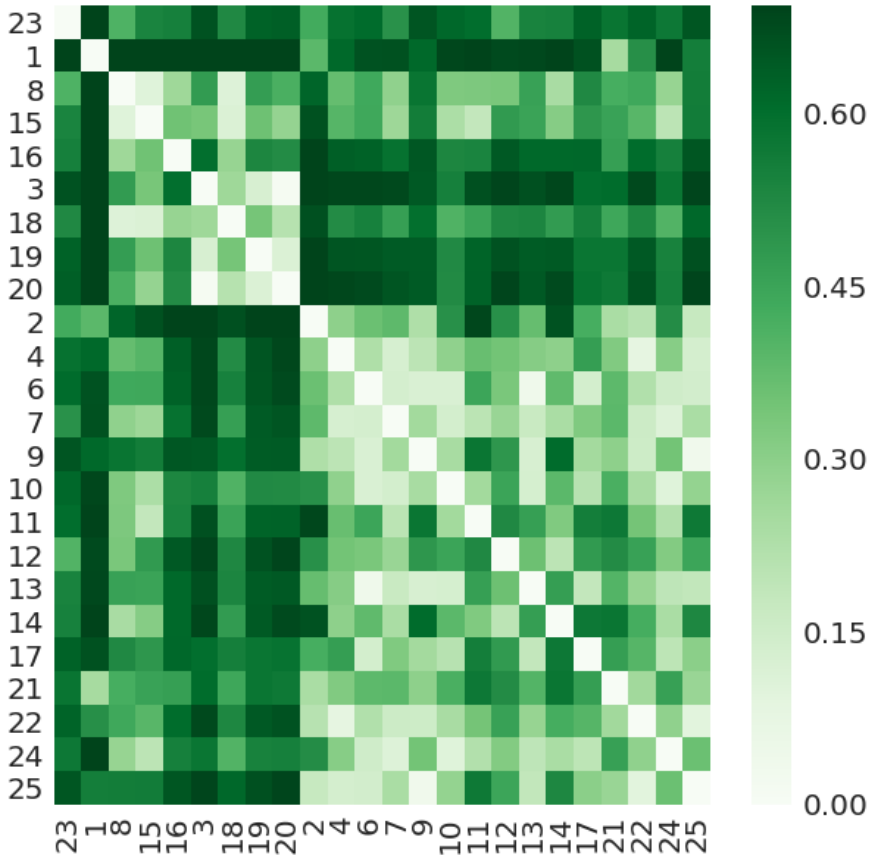


FIGURE 3.33: JS divergence for Tyr-ATP distance distributions. Clustering described in section 3.2.8, using ϵ=0.15 with 5 states.

All distance analysis has also been completed for a set of two "swapped ligand" simulations, where activator JS30 and inhibitor 1F8 were modelled into structures 3ORX and 3OTU respectively. Results are found later in section 3.3.7.

3.3.3 Torsion KL

The distances previously calculated rely on some structural knowledge of the system under study. However it is useful to ensure that analysis carried out is not biased, and therefore require methods which do not rely on any prior knowledge. The first method used to carry out a more unbiased analysis is to compute distributions of all torsional angles, and use KL divergence to compare between the activated and inhibited systems.

3.3.3.1 KL testing

Testing was carried out by changing the prior count added to each bin. As mentioned in section 2.3.1, equation 2.19 describes addition of a uniform prior count to all bins. The factor of x determines the value used for the prior count, as a fraction of the mean value is added into each bin before distributing the data. Values of x of $\frac{1}{10}$, $\frac{1}{10^2}$, $\frac{1}{10^3}$, $\frac{1}{10^4}$, $\frac{1}{10^5}$ and $\frac{1}{10^6}$ were tried initially. For two distributions which are almost continuous (for example, from two different simulations of the same system), the KL value obtained converges using a value of x of $\frac{1}{10^3}$, or smaller. However in the case of a poor overlap of distributions, this has a more pronounced effect, and difficulties could therefore arise if comparing two KL values; where one is obtained for two well overlapping distributions and the other is for two poorly overlapping distributions, and so the result is not consistent. There are various methods which can be used to solve this. In this case, a fixed value was split over all zero-valued bins only, rather than a uniform amount being added to all bins. In this way, the amount added to an empty bin is effectively weighted depending on the extent of the overlap between two distributions.

Further testing was completed to determine if the KL values obtained were due to differences between the systems, or an artefact of finite sampling. To achieve this, a single trajectory was split into 4 segments, and the KL divergence computed between the different segments of the same simulation. For each trial completed in this way, the KL values were negligible when compared to the KL values obtained by comparing two systems.

3.3.3.2 KL on original compound set

Analysis is carried out by computing distributions of ψ, ϕ, $\chi 1$ and $\chi 2$ angles. For each angle, this is plotted as a normalised histogram for four different systems: PDK1 only, PDK1 with activating ligand JS30 (3OTU), PDK1 with activating ligand 2A2 (3ORZ) and PDK1 with inhibiting ligand 1F8 (3ORX).

The KL values obtained were then visualised as Cα "B-factor" using the software PyMol [99]. This allowed simple identification of areas where torsional angles differed the most between activated and inhibited systems and results are also shown for summed "backbone" ($\psi + \phi$) and "sidechain" ($\chi 1 + \chi 2$) KL values. The largest differences in backbone torsions between activating ligand JS30 (PDB ID 3OTU) and inhibiting ligand 1F8 (PDB ID 3ORX) highlighted several residues around the hinge of the activation loop (3.34). Results are similar when comparing another activator 2A2 (PDB ID 3ORZ) to the same inhibitor 1F8 (PDB ID 3ORX) as shown in figure 3.35. Differences in side chains were closer to the ATP binding site and included residues in the functionally relevant DFG-loop. Again, similar regions are highlighted when comparing activating ligand 2A2 (3ORZ) to the same inhibitor.

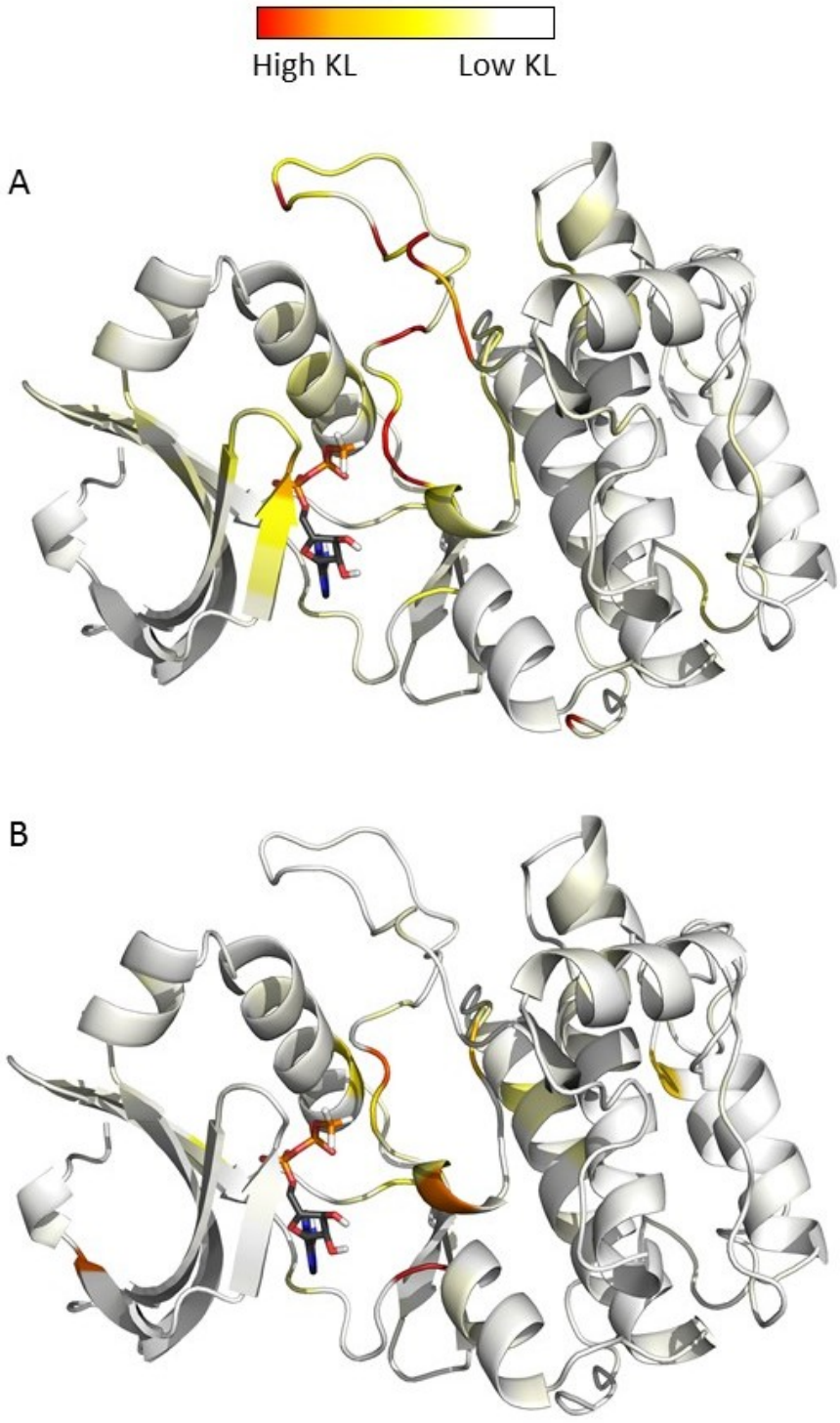


FIGURE 3.34: KL-divergence for A: backbone and B: sidechain torsional angles for $KL(JS30|1F8)$.

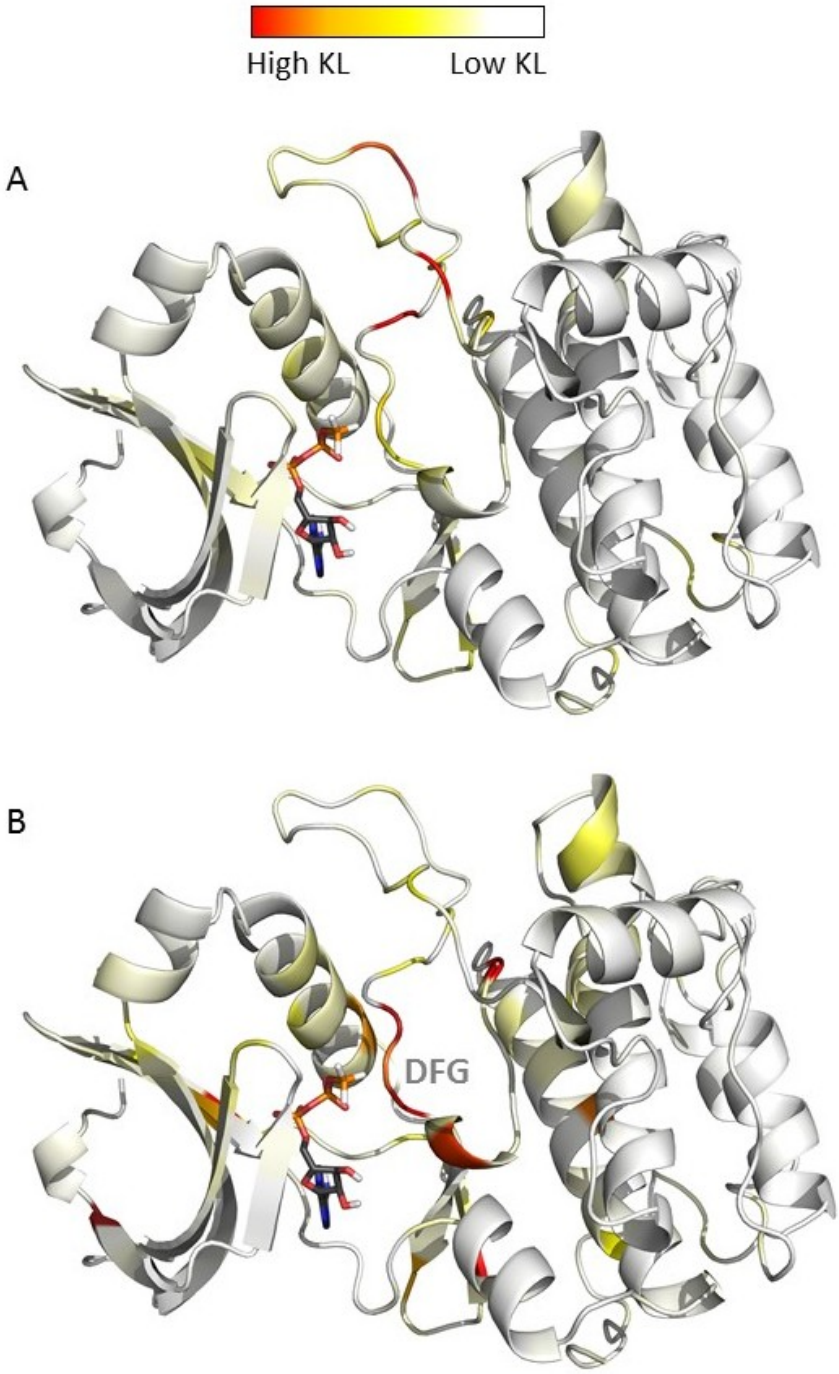


FIGURE 3.35: KL-divergence for A: backbone and B: sidechain torsional angles for $KL(2A2|1F8)$.

3.3.4 PCA on Cα coordinates

3.3.4.1 Initial compound set

Initial analysis included only the four simulations from Set A as in table 3.2: ligands 1F8, 2A2, JS30 and with no allosteric ligand. The first and second eigenvectors correspond to 26 % and 20 % of the variance respectively. PC1 coresponds to the activation loop motion, which contributes to the largest percentage of variance due to differences between activated and inhibited complexes. Per atom contributions to PC1 were visualised as per figure 3.36A. The highest atom contribution to PC1 is the residues of the activation loop (residues Ser159-Asn168).

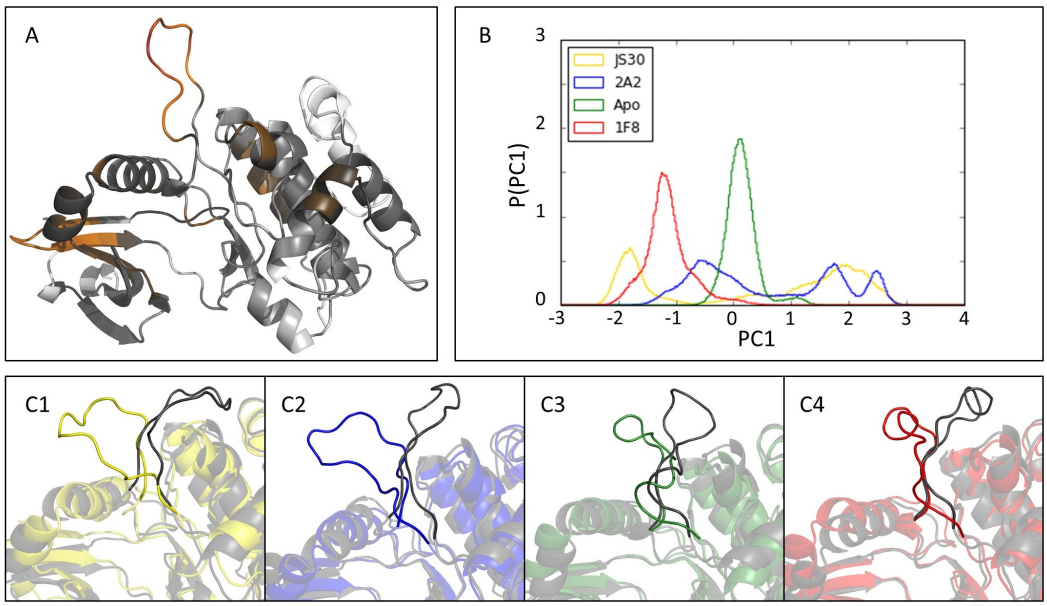


FIGURE 3.36: PC1 of four simulations: Yellow: activator JS30; Blue: activator 2A2; Green: apo; Red: inhibitor 1F8. A: Per residue contribution to PC1. Colour scheme is white-grey-red with increasing contribution. B: Distributions of PC1 for each system. C1-C4: structures corresponding to minimum (grey) and maximum (colour) values of PC1.

With either of the two activating compounds bound (JS30 and 2A2), this loop visits a wider range of conformations than with inhibitor bound (1F8), this can be seen by considering the distributions in figure 3.36B. Structures

representing the maximum and minimum values for PC1 for each system are shown in figure 3.36C.

PC1 results are in line with previously discussed Kullback-Leibler (KL) diversion results in section 3.3.3 between activated and inhibited systems using backbone torsional angles, which also highlighted that the highest differences were found in a hinge region of the activation loop (figures 3.34A and 3.35A).

As discussed in section 3.3.2.3, a notable difference between activator and inhibitor bound simulations is in the conformation of Tyr54. The distance from Tyr54 to ATP for both activators, and the apo simulations is shorter, but they also have a wider range of values than for inhibitor bound. This could affect the conformation of the activation loop. In the inhibitor bound simulation, Tyr54 is flipped away from ATP, and makes an interaction with the phosphoserine on the activation loop (figure 3.29). In the activator bound simulations, Tyr54 adpots both conformations: towards ATP, and flipped away. In the inhibitor bound this interaction seems more stable, which could restrict the motion of the loop. The more dynamic conformation of Tyr54 in the activator bound simulations could allow for the loop conformations corresponding to the maximum PC1 values, which are shown in figure 3.36.

The result is that in simulations with allosteric inhibitor or with no allosteric ligand, the variance of this loop motion is substantially less than for either of the two activators. Distributions of these values highlight this, with only the two activators having values at the higher range of PC1 (figure 3.36B).

The results for PC2 show that *all* allosteric ligand bound simulations have a shift in the distribution of PC2 relative to apo, as shown in figure 3.37B, but there doesn't seem to be similarity between the two activators. Per atom contributions to PC2 show that this predominantly involves the residues of helix-B, and β strands 1 and 2.

Plots of PC1 and PC2 values per snapshot, and a two-dimensional plot of PC1 and PC2 with each trajectory overlaid, can be found in the appendix

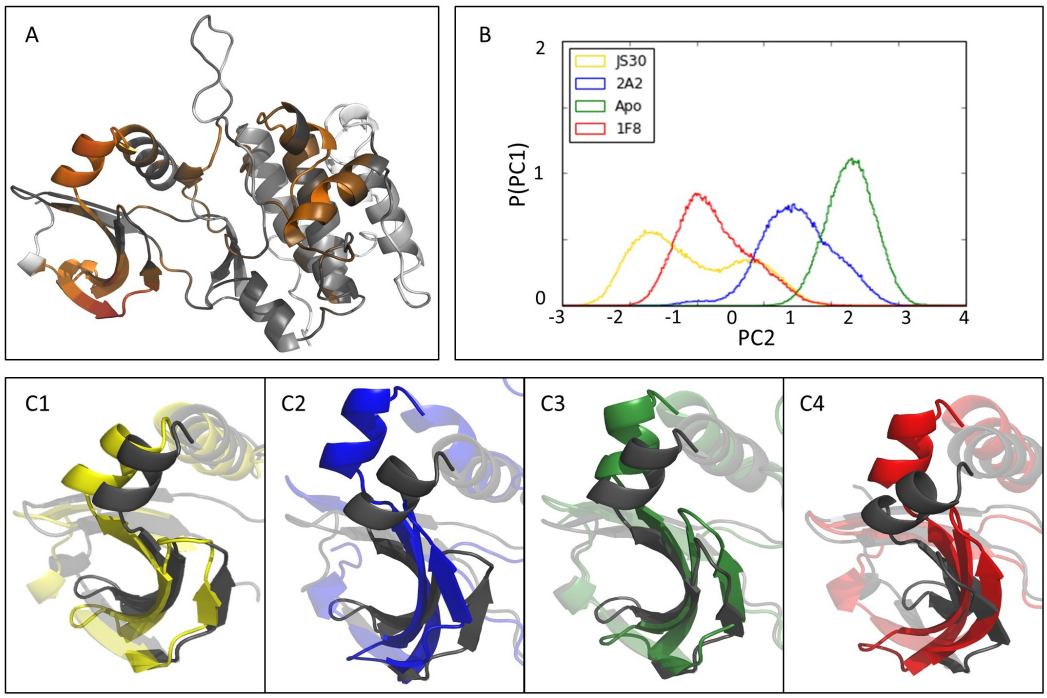


FIGURE 3.37: PC2 of four simulations: Yellow: activator JS30; Blue: activator 2A2; Green: apo; Red: inhibitor 1F8. A: Per atom contribution to PC2. B: Distributions of PC2 for each system. C1-C4: structures corresponding to minimum (grey) and maximum (colour) values of PC2.

3.3.4.2 Extended compound set

PCA was then extended to include further ligands from the set for which activity data is provided by Sadowsky et al. [91]. Compounds were selected to cover a range of activities, from both scaffold A and scaffold B (which are the scaffolds for the two crystal structures 3ORZ and 3OTU). The compounds selected can be seen in table 3.2, and this analysis includes set A and B from this table.

Again this highlighted activation loop motion as the first principal component, which describes 26% of the variance in the dataset. Compounds were then separated based on activity, and this highlighted the differences between the inhibitor bound and activator bound simulations. Only the highest activating compounds shown in figure 3.38(C) show values for PC1 higher than 1.5. These values of PC1 correspond the the conformation of the

loop shown in 3.38(C) in grey, where the activation loop moves much closer to alpha helix C. Based on this subset of data, it appeared that the highest activating compounds all share a conformation not accessed by the lower activating, or inhibiting ligand bound simulations. As the helix C is located between the allosteric and active sites, and the loop interacts with this helix at Tyr54, is is possible that the allosteric ligands affect the helix C, which in turn affects the loop position, and this has some influence on the activity. The position of the loop can affect the substrate binding, as this binds near to the the γ-phosphate of ATP, which is near the hinge region of the loop as seen in figure 3.11.

From these distributions, the structures corresponding to the positive values of PC1 which show the highest population for each system were output, and can be seen in figure 3.39. This shows that the loop is closer to α-helix C in both cases. For the inhibitor and apo simulations, the loop is much further from α-helix C.

It can then be seen that the α-helix C position varies in the structures where the activation loop is closer to α-helix C. A closer look at the position of this helix shows that in the apo and inhibitor bound conformations, the helix is much closer to the allosteric site. However when inhibitor is bound, this moves closer to the active site, as in figure 3.40.

This helix contains Glu58, previously discussed in section 3.3.2.2, which forms a salt bridge with Lys39. Comparing the inhibited structure to the two highest activators, it seems that the interactions with ATP are increased as the helix moves closer to the active site, with the activating compounds, shown in figure 3.41.

In order to attempt to validate this trend, JS divergence values were computed using the distributions of PC1 for this compound set. The results were clustered, and summarised in figure 3.42.

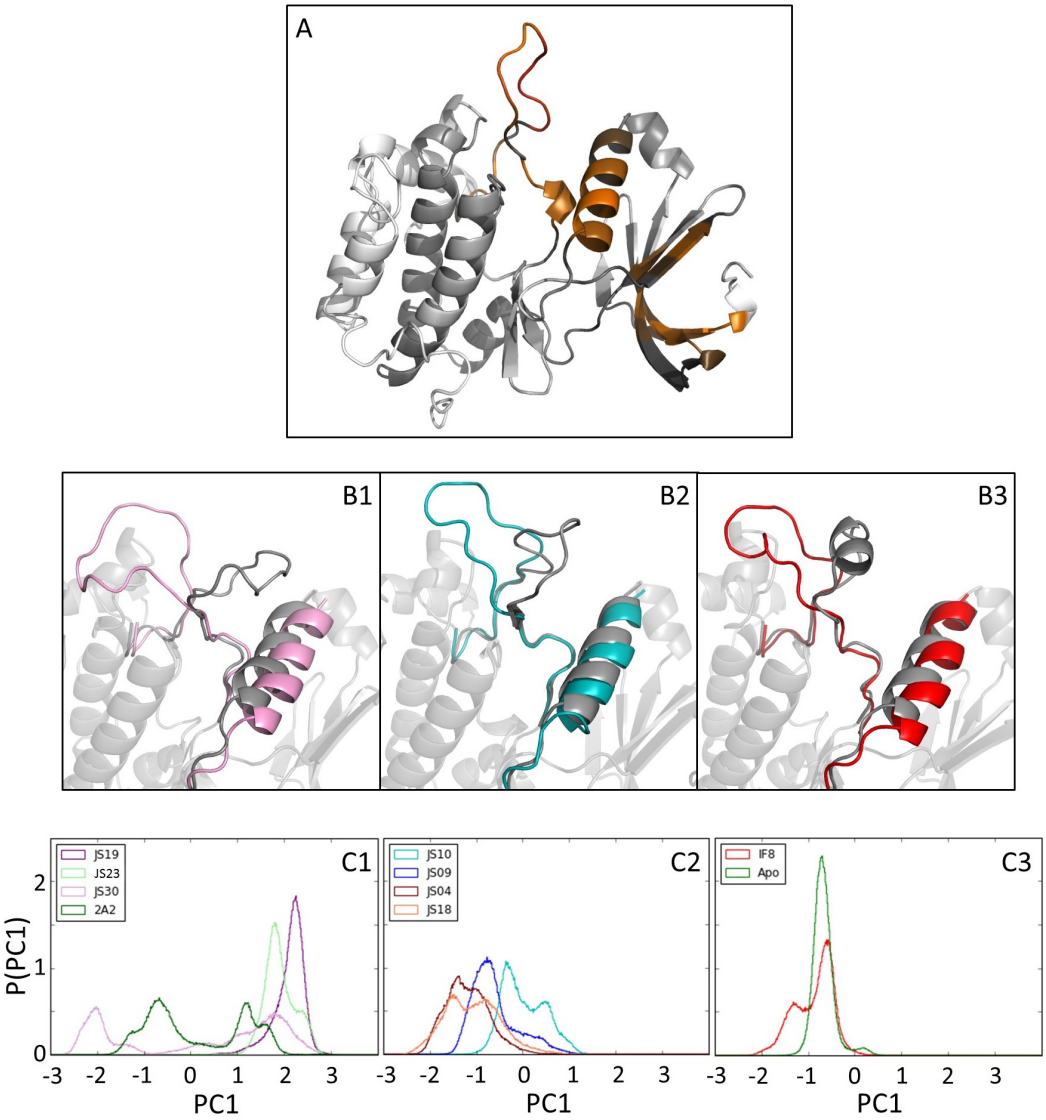


FIGURE 3.38: PC1 for extended compound set. A: Per atom contribution to PC1. Colour scale White-Grey-Orange-Red with increasing KL value. B: structures representing maximum and minimum PC1 values for B1: JS30; B2: JS10; and B3: 1F8. C: Distributions for PC1. Compounds separated based on activity with C1: highest activators; C2: medium activators; and C3: apo and inhibitor.

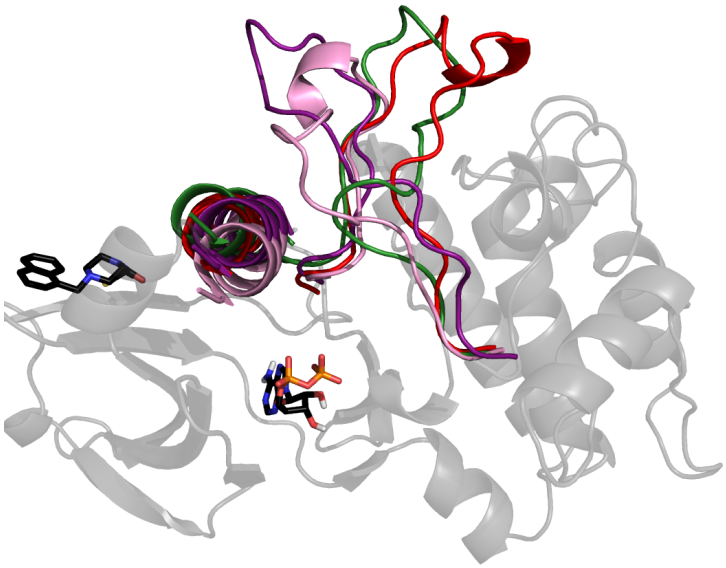


FIGURE 3.39: Conformations representing the high values of PC1 at the most populated point of each distribution for JS19 (purple) JS30 (lilac) 1F8 (red) and apo (green). Highlighted regions are the activation loop and helix C. The allosteric site is on the left, with ATP bound at the central active site, and both are shown in sticks.

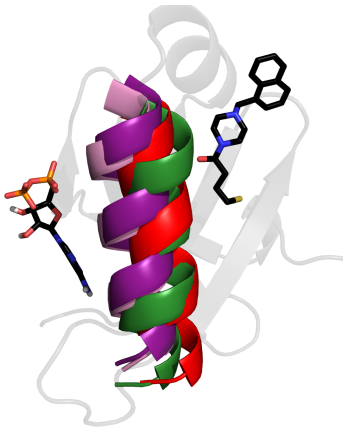


FIGURE 3.40: Helix C for JS19 (purple), JS30 (lilac), 1F8 (red) and apo (green). Allosteric ligand JS30 is shown on the right, and ATP bound at the active site is shown on the left.

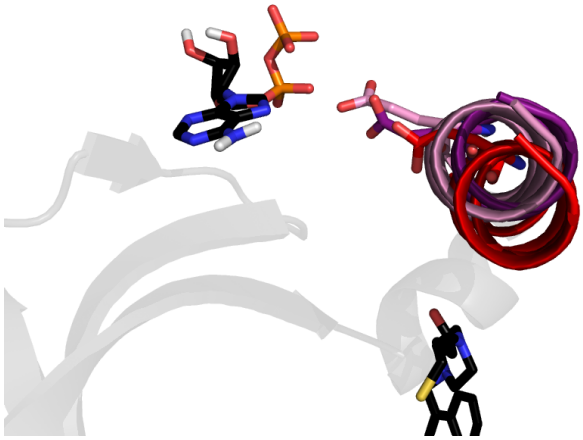


FIGURE 3.41: Helix C with Glu58 for JS19(510-purple), JS30(630-lilac) and 1F8(inhibitor-red). Allosteric ligand JS30 is shown below helix C, and ATP bound at the active site is shown on the top left.

The second principal component highlights many residues around the allosteric site, in particular the smaller α-helix B adjacent to α-helix C. The difference in position of helix B, could affect the position of helix C, and so could also contribute to the difference between the active and inhibited conformation. As compounds based on scaffold B are longer than scaffold A due to the extra CH_2, they sit further towards this smaller helix and could cause this shift.

Therefore, distributions for PC2 were plotted, and separated based on scaffold. With the exception of compound JS09, there seems to be separation of PC2 based on scaffold. JS09 could be an outlier, as the structure has a fused ring to R2 and R3 (see figure 3.16 and table 3.1), which points directly towards this smaller helix (α-B).

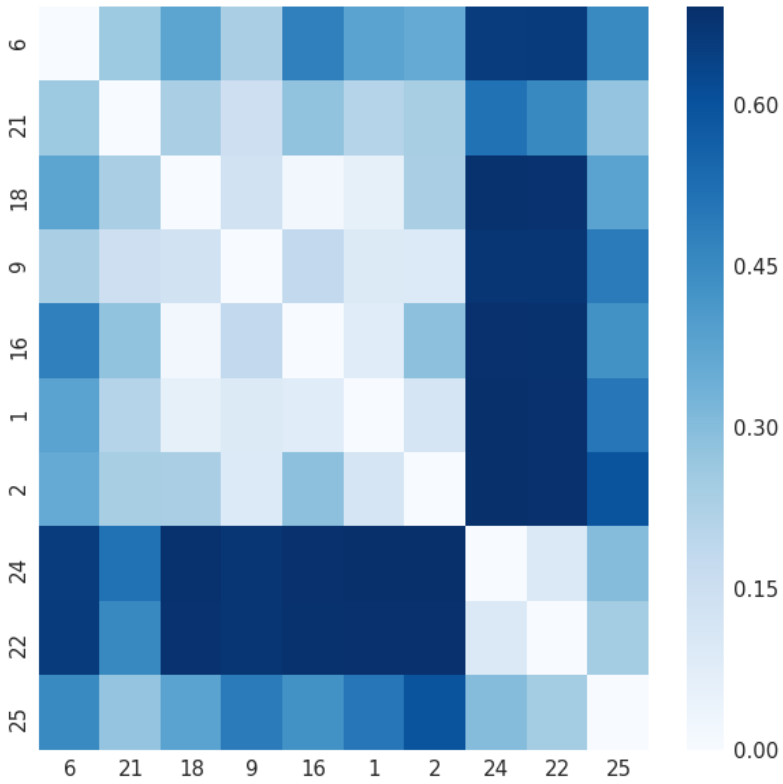


FIGURE 3.42: JS divergence for extended compounds set of PC1 distributions. Clustering described in section 3.2.8, using ϵ=0.2 with 3 states.

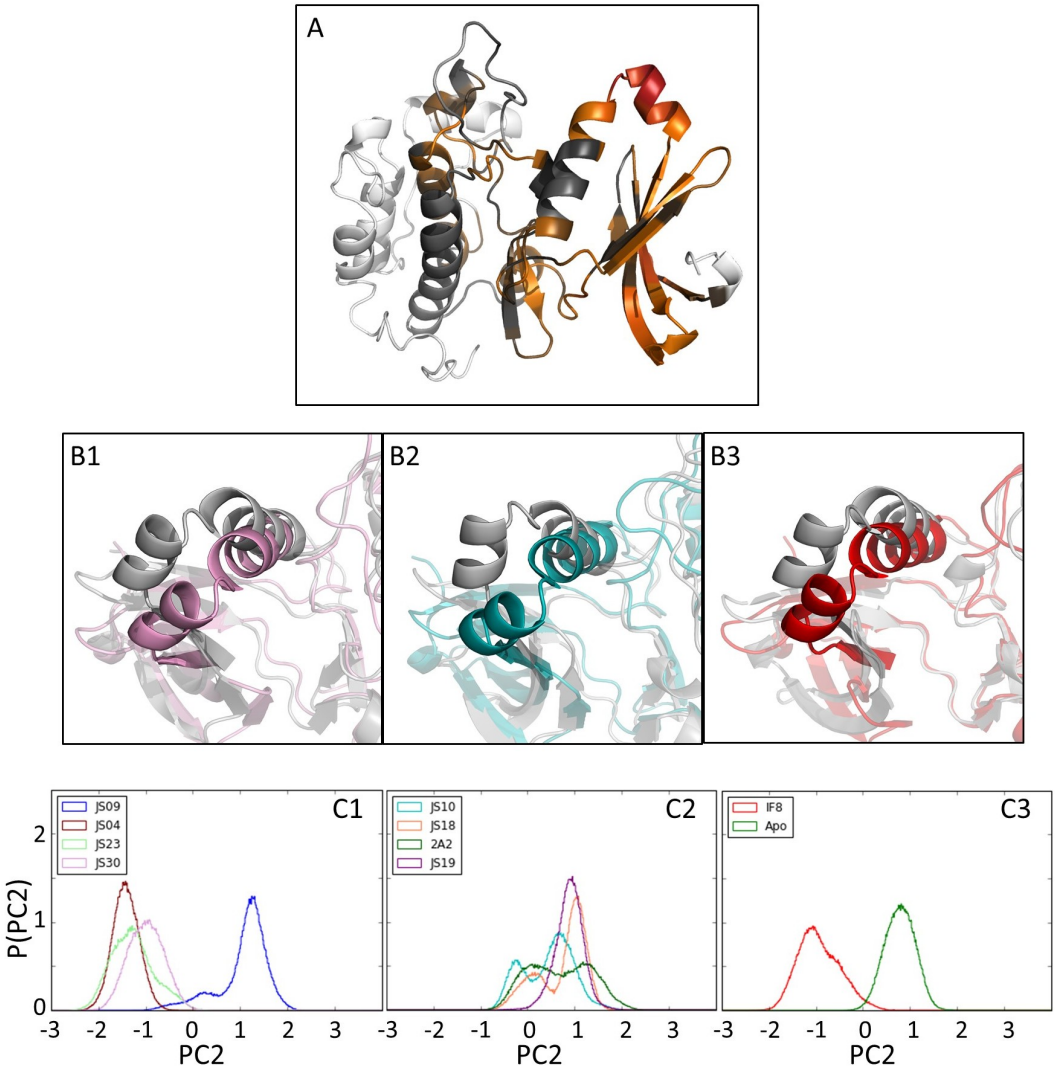


FIGURE 3.43: PC2 for extended compound set. A: Per atom contribution to PC2. Colour scale White-Grey-Orange-Red with increasing KL value. B: structures representing maximum and minimum PC2 values for B1: JS30; B2: JS10; and B3: 1F8. C: Distributions for PC2. Compounds separated based on scaffold with C1: scaffold B; C2: scaffold A; and C3: apo and inhibitor.

3.3.4.3 Full compound set for scaffold A and B

The full set of compounds based on scaffold A and B can be seen in table 3.2, and PCA results discussed so far only included a subset of these compounds. In order to attempt to validate the results, PCA was run for the full set of simulations. The first and second principal components are very similar to that obtained for the previous sets of compounds. However, now the full set of compounds has many more compounds of similar activities. The results highlight that the trend noticed for the previous set of compounds does not extend to the full set. The inhibitor still shows different behaviour to any of the activating set, however it is not possible to determine the degree of activation as some of the lower activating compounds show the same PC1 distributions as high activating compounds. To highlight any differences, the JS divergence of PC1 was clustered as shown in figure 3.44.

Results for PC2 also are not entirely clear however do show some separation of scaffold A and B compounds.

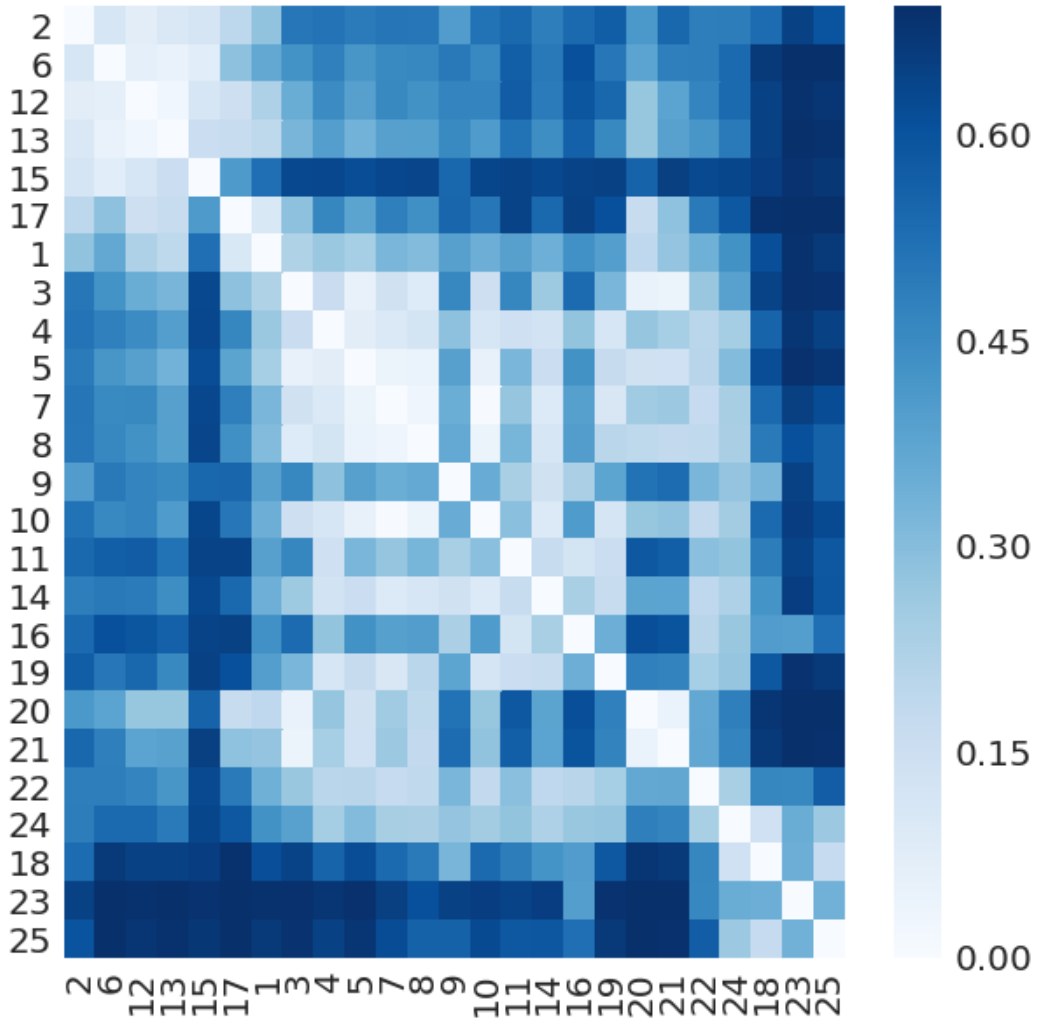


FIGURE 3.44: JS divergence for full compound set on PC1. Compounds numbered as in table 3.2. Clustering described in section 3.2.8, using ϵ=0.07 with 3 states.

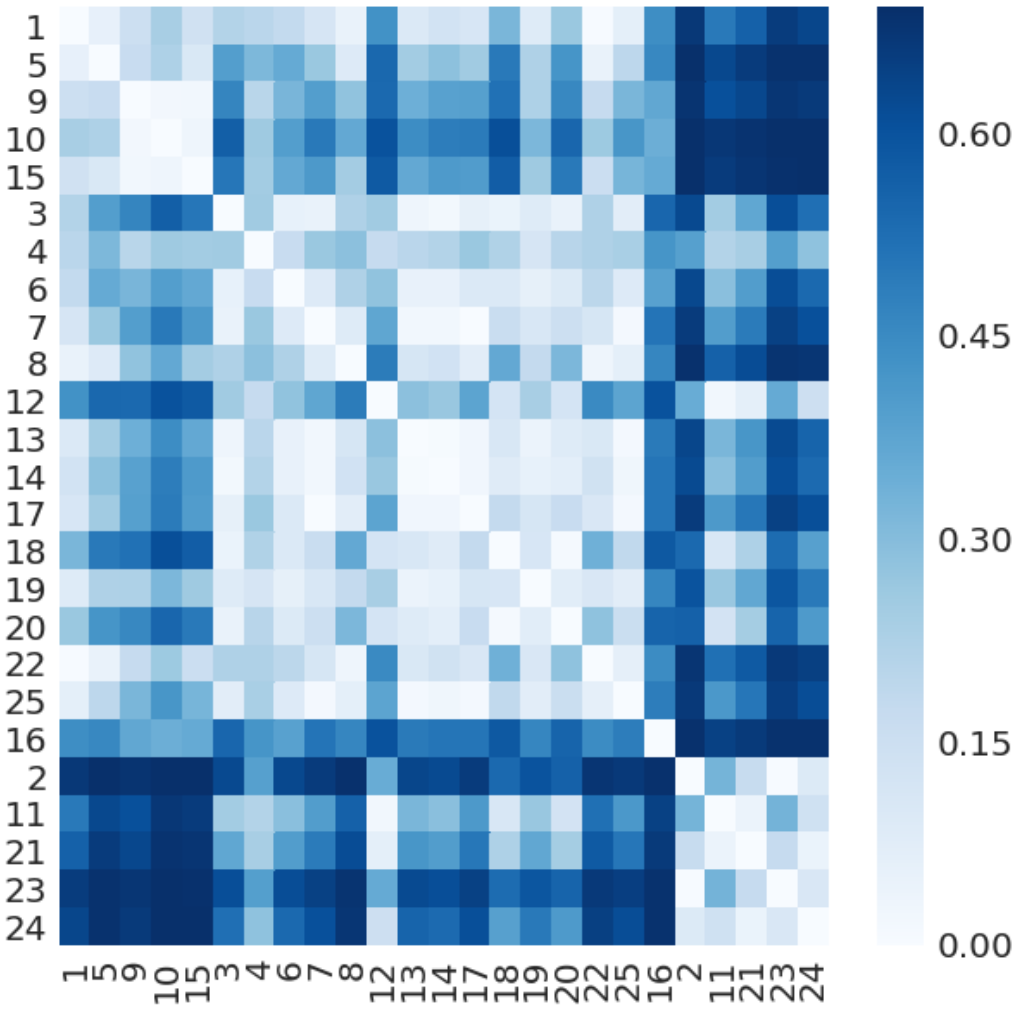


FIGURE 3.45: JS divergence for full compound set on PC2. Compounds numbered as in table 3.2. Clustering described in section 3.2.8, using ϵ=0.05 with 4 states.

3.3.5 Energy decomposition

The geometric descriptors previously discussed are useful in determining differences between activated and inhibited conformations, however this does not directly confirm trends in activity. Many of these measurements are based on information which can be obtained from the crystal structure, or knowledge of the system under study, and so require advance knowledge of the differences between active and inhibited complexes. Other differences (e.g. PCA) show structural differences but do not confirm if these are important for activity as they cannot directly correlate the high variance motions with a functionally relevant descriptor. One potential way to highlight important structural changes without this bias, is to compute interaction energies between various residues. This allows easier identification of which structural changes could be related to changes in activity. This can be achieved by decomposing the potential energy surface into per residue interactions, allowing identification of which residues are responsible for binding, or alternatively which are blocking access to ligand/substrate via repulsive interactions. Also important to understand is the structure-activity relationship of the ligands bound to the allosteric site. By computing interactions of the allosteric ligand with the protein, it may be possible to explain differences in activity. Energy analysis was computed for various residues, and also for all residues in the peptide which could then be summed to give overall interaction energies of the peptide with the protein/ATP. All energies quoted are a sum of the Coulombic and Lennard Jones interaction energies.

3.3.5.1 Peptide interactions

Interactions of the peptide were computed, in order to attempt to highlight any differences that may affect substrate access to ATP. Summing all protein residue interactions with the peptide show that for both activator bound and inhibitor bound simulations, the overall peptide-protein interactions are attractive and show no particular trend. Results can be seen in table 3.7. The differences in how the peptide interacts when comparing activators to

Ligand Ligand	Peptide Protein	PepThr Protein	ATP Pep-Thr	ATP Entire peptide
1JS10	-120.71	-3.59	-16.36	-168.92
2JS18	-97.83	-7.9	-21.83	-184.22
3JS19	-133.74	-13.96	-32.17	-187.62
1JS09	-100.62	-13.54	-24.02	-233.65
2JS04	-138.96	-10.02	-28.08	-161.29
3JS23	-104.38	-7.94	-25.02	-230.19
1F8	-122.5	-23.9	2.94	-214.65
APO	-118.1	-0.93	-39.45	-215.8
2A2	-125.84	-12.25	-24.44	36.11
JS30	-23.7	-0.68	-20.24	-158.91

TABLE 3.7: Interaction energies of the substrate peptide with different parts of the system. Peptide-protein interactions; peptide Thr residue with protein interactions; ATP with peptide Thr; and ATP with the entire peptide.
Energies in kcal mol^{-1}

.

the inhibitor is only in the peptide Thr residue and the interactions of this residue with ATP.

To further understand how the interactions vary, the interaction of the peptide with each residue of the protein was input as a B-factor and visualised, where residues highlighted in blue are attractive interactions and residues shown in red are repulsive, as shown in figure 3.46. This shows that with inhibitor 1F8 bound, the interactions around ATP do seem to vary when compared to activator 2A2; and that particularly the repulsive interaction of Lys39 could result in differences in positioning of the substrate.

3.3.5.2 Allosteric ligand

Interactions of the allosteric ligand with the protein could help to identify the differences in ligand structure which lead to differences in function. The most notable difference in interactions between activating ligand 2A2 (2ORZ) and inhibiting ligand 1F8 (3ORX) was with Arg59 (figure 3.47).

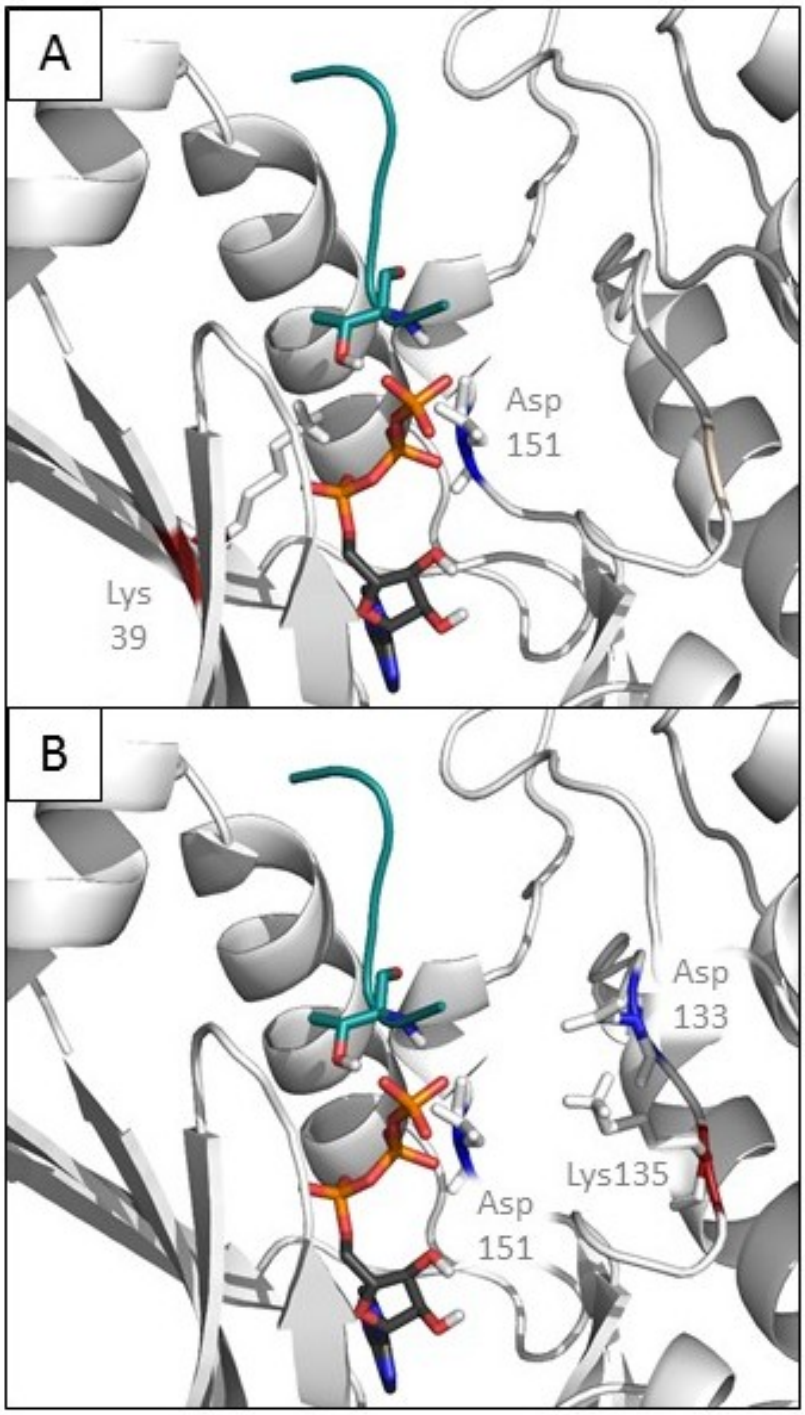


FIGURE 3.46: Interaction energies of the peptide with protein residues for A: inhibitor 1F8; B: activator 2A2; and C: activator JS30.

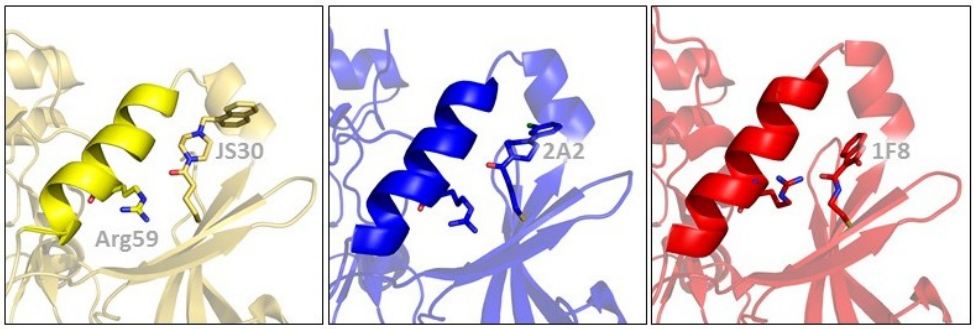


FIGURE 3.47: Conformation of Arg59 in crystal structures 3OTU (yellow), 3ORZ (blue) and 3ORX (red).

It can be seen from the crystal structure that the side chain adopts different conformations in these PDK1-inhibitor and PDK1-activator complexes. Analysis of the energy interactions suggests that with inhibitor bound, there are larger attractive interactions with this residue and potentially this could result in the rotation of helix αC. Adjacent to Arg59, Glu58 forms the salt bridge with Lys39 as previously discussed, and disruption of this interaction changes the interactions with ATP. Therefore if changes in interactions with Arg59 cause shifts in this helix, this could explain the changes in interactions with ATP.

This residue is also highlighted when calculating the KL divergence of the interaction energies. Figure 3.48 highlights high KL values in red for activator 2A2 (3ORZ), calculated relative to inhibitor 1F8 (3ORX). However activator JS30 (3OTU) does not show this same pattern of interactions yet still results in shift of helix C in a similar conformation to activator 2A2. Since JS30 is a bulkier ligand and so taking up more space in the allosteric pocket, and also shows differences in the interaction energies on the helix adjacent to C, it is possible that helix C can still adopt the active conformation but without the same interaction pattern. The conformation of the helix adjacent to helix C could affect the position of helix C, and this could be sufficient for the change in activity.

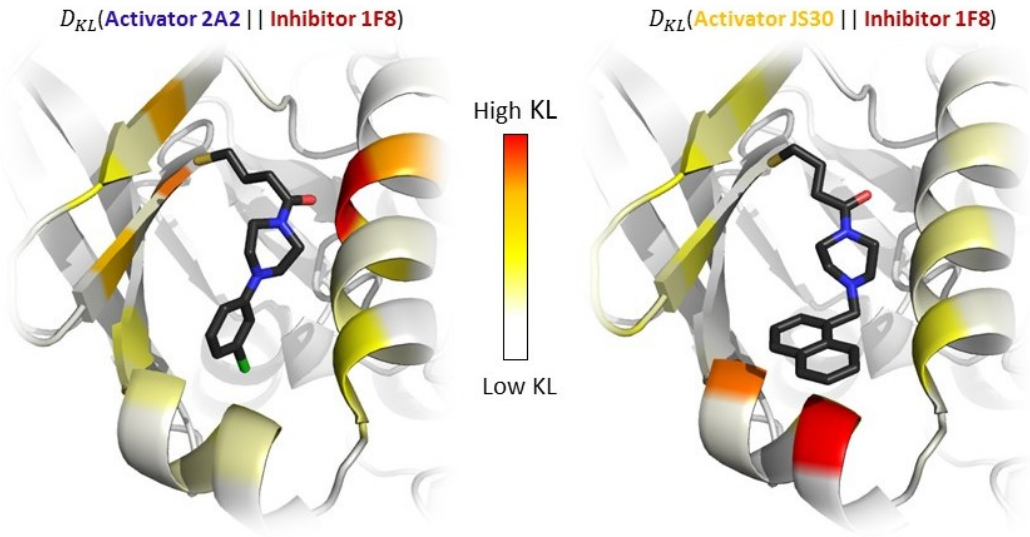


FIGURE 3.48: KL divergence of interaction energies computed between activator 2A2 and inhibitor 1F8, and between activator JS30 and inhibitor 1F8.

3.3.5.3 ATP

Interactions of ATP were computed. As discussed, there is a difference when inhibitor is bound, in the distance between two residues Lys and Glu which form a salt bridge adjacent to the ATP binding site. Interaction energies of these two residues with ATP vary between activated and inhibited complexes (Lys39: Activator JS30 -79.32 $\pm$ 0.72 kcal mol^{-1}. Inhibitor 1F8 -14.31 $\pm$ 4.33 kcal mol^{-1}. Glu58: Activator JS30 62.61 $\pm$ 0.8 kcal mol^{-1}. Inhibitor 1F8 3.62 $\pm$ 0.82 kcal mol^{-1}). These changes in interactions appear to shift the position of ATP in the active site. Overall interactions of ATP with the protein were determined by summing all individual interactions and indicate that activator bound PDK1 stabilises ATP, which could also promote reactivity. For both the activated and inhibited complexes, the interactions of ATP with the substrate peptide seem reasonably similar, however then considering the interaction of ATP with only the threonine which the phosphate is transferred to, it can be seen that the inhibitor bound simulation has repulsive interactions (2.94 $\pm$ 2.15 kcal mol^{-1}, while all other simulations have attractive interactions (between -16 and -40 kcal mol^{-1}). The values for the extended set (including sets A and B from 3.2) can be seen in table 3.8.

Ligand	Activity	A	B	C	D	E	F	G
1F8	50	169.02	-214.65	-45.63	-14.31	3.62	-10.69	2.94
Apo	100	19.07	-215.80	-196.73	-14.71	4.65	-10.06	-39.45
JS10	210	74.8	-168.92	-94.12	-80.71	56.67	-24.04	-16.36
JS09	240	-3.39	-233.65	-237.04	-139.83	35.41	-104.42	-24.02
JS04	330	26.1	-161.26	-135.16	-92.84	59.75	-33.09	-28.08
JS18	370	-24.76	-184.22	-208.98	-120.47	80.33	-40.14	-21.83
2A2	394	100.21	-183.37	-83.16	-81.35	63.05	-18.3	-24.44
JS23	460	106.66	-230.19	-123.53	-59.69	19.85	-39.84	-25.02
JS19	510	73.09	-187.63	-114.54	-88.49	37.38	-51.11	-32.17
JS30	630	79.68	-158.91	-79.23	-79.32	62.61	-16.71	-20.24

TABLE 3.8: Energies of interactions of ATP with various other parts of the system. Activity shown is as a percentage relative to Apo (where Apo is 100%). A: ATP interactions with the entire protein. B: ATP interactions with the substrate peptide. C: ATP interactions with the protein and substrate combined. D: ATP interactions with Lys39. E: ATP interactions with Glu58. F: ATP interactions with Glu58 and Lys39 combined. G: ATP interactions with the substrate Thr. Energies in kcal mol^{-1}.

3.3.6 Mutual information

3.3.6.1 MI testing

To confirm that MI results were reasonable, a test dataset was constructed consisting of three different distance measurements: Asp78 to Gln162 (A); Asp78 to Ser160 (B); and Ile100 to Asp142 (C). Asp78 is located in the C-terminal lobe, while Gln162 and Ser160 are part of the activation loop. Therefore, MI should be high for the A-B pair, as both are related to the same loop motion and so should be correlated. Distance C was computed between two residues located on the N-terminal lobe. This distance is not related to loop motion, it is expected to show no correlation to the other two distances (figure 3.49). MI was computed for each distance using the simulation with no allosteric ligand, and results can be found in table 3.9.

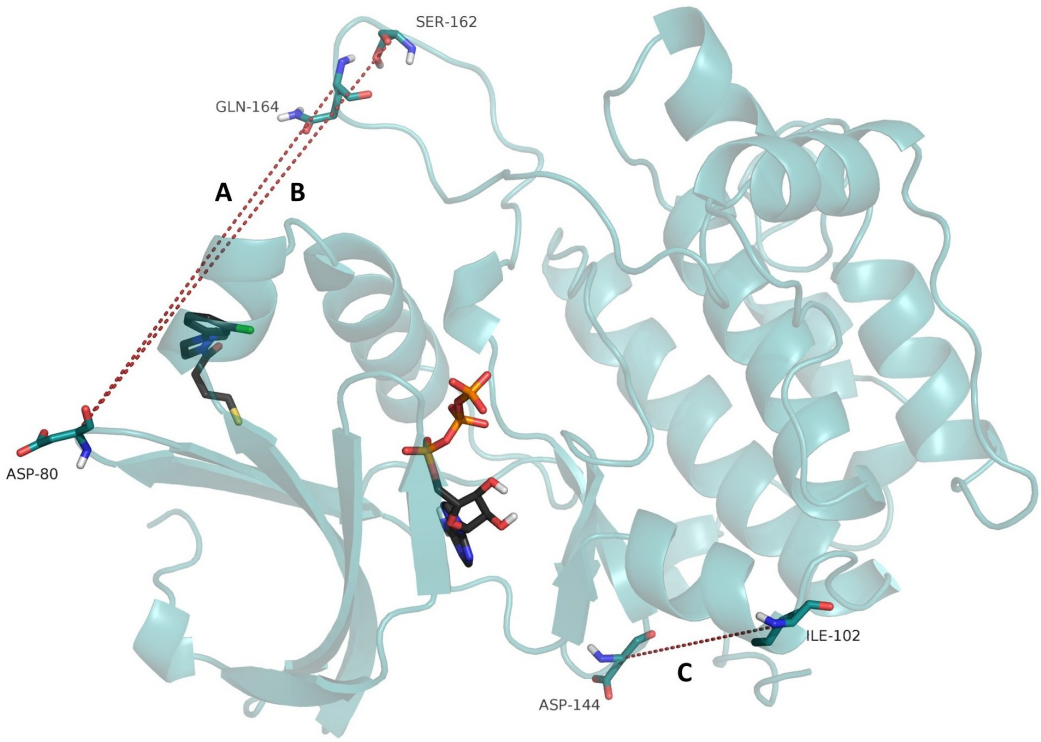


FIGURE 3.49: Distances computed to check MI results are reasonable. Distances A and B should be correlated, while A,C and B,C should show less correlation.

I(A;B)	*I(A;C)*	*I(B;C)*
0.985	0.134	0.142

TABLE 3.9: MI computed between distances A, B and C. Distances A, B and C are highlighted in 3.49

Values for $I(A;B)$ are high, which suggests they show correlation, while both $I(A;C)$ and $I(B;C)$ have low values, suggesting no correlation.

It is important to ascertain whether a particular MI signal is true, or a result of noise due to finite sampling. To establish this, MI was first computed between two sets of data. One set was then randomised in time, and MI computed again. This randomised MI is then subtracted from the computed MI to give I_{corr}. This MI was calculated for the same distances shown above, and extended to include three further simulations. The results can be seen in table 3.10.

	$I(A;B)_{corr}$	$I(A;C)_{corr}$	$I(B;C)_{corr}$
APO	0.872	0.020	0.028
1F8	0.833	0.019	0.031
2A2	1.184	0.007	0.012
JS30	1.069	0.075	0.087

TABLE 3.10: MI computed between distances A, B and C using in house script, for four systems. Values reported as MI_{corr} using 200,000 snapshots and 300 bins.

This was then computed for a range of bin numbers to establish how both the original MI, and the randomised MI, vary depending on binning. Taking the difference of both MI (unedited data, and then with one dataset randomised) for each number of bins used (between 1 and 1000 bins), it was possible to determine a maximum value of MI which can be attributed to a real signal and not a result of noise (figure 3.50). The number of bins corresponding to the largest difference between these two data sets was used for all further calculations. To facilitate analysis, MI was initially computed on a reduced trajectory composed of 1 every 5 snapshots, reducing the data points to 40k snapshots. In this case, 60 bins were used. Results for each

system are shown in figure 3.50 and show the MI computed between the interaction energy of ATP with the protein, and the first principal component. MI for a range of bin numbers is higher in the unedited data compared to the randomised set in all ligand bound simulations. However, no difference in MI is seen between original and randomised with apo simulations, implying that for this simulation no correlation can be found between ATP interaction energy and PC1.

These plots indicated that for the dataset comprising of 40,000 snapshots, that around 60 bins maximised the MI signal relative to noise. Therefore the following MI results presented are for MI_{corr}, as defined in 2.3.3, using 60 bins for datasets composed of 40,000 snapshots, and 100 bins for datasets composed of 100,000 snapshots.

Further data relating to testing of MI can be found in appendix A.

3.3.6.2 MI results: Original compound set

As previous analysis suggests that the hydrogen bonding with ATP may play a role in the changes of activity, it would be useful to determine if the motion described by PC1 is related to the stability of ATP in the active site. MI was therefore calculated between the value of PC1, and the interaction energy of ATP with the protein.

As distances A and B are also related to the loop motion, MI was first computed for the three distances as defined in 3.49, to confirm that distances A and B show different MI to PC1, than distance C. In all cases, when there is allosteric ligand bound, MI for both distances A and B with PC1 is higher than for distance C for PC1. In all cases for the apo simulation, MI between any of the distances with ATP interaction energies is negligible (table 3.11).

MI was then computed between ATP interaction energy, and value of PC1 for each system. Results can be seen in table 3.12. To facilitate analysis, only every second snapshot was used (100k snapshots) for a 1 μs trajectory, and is calculated using 100 bins.

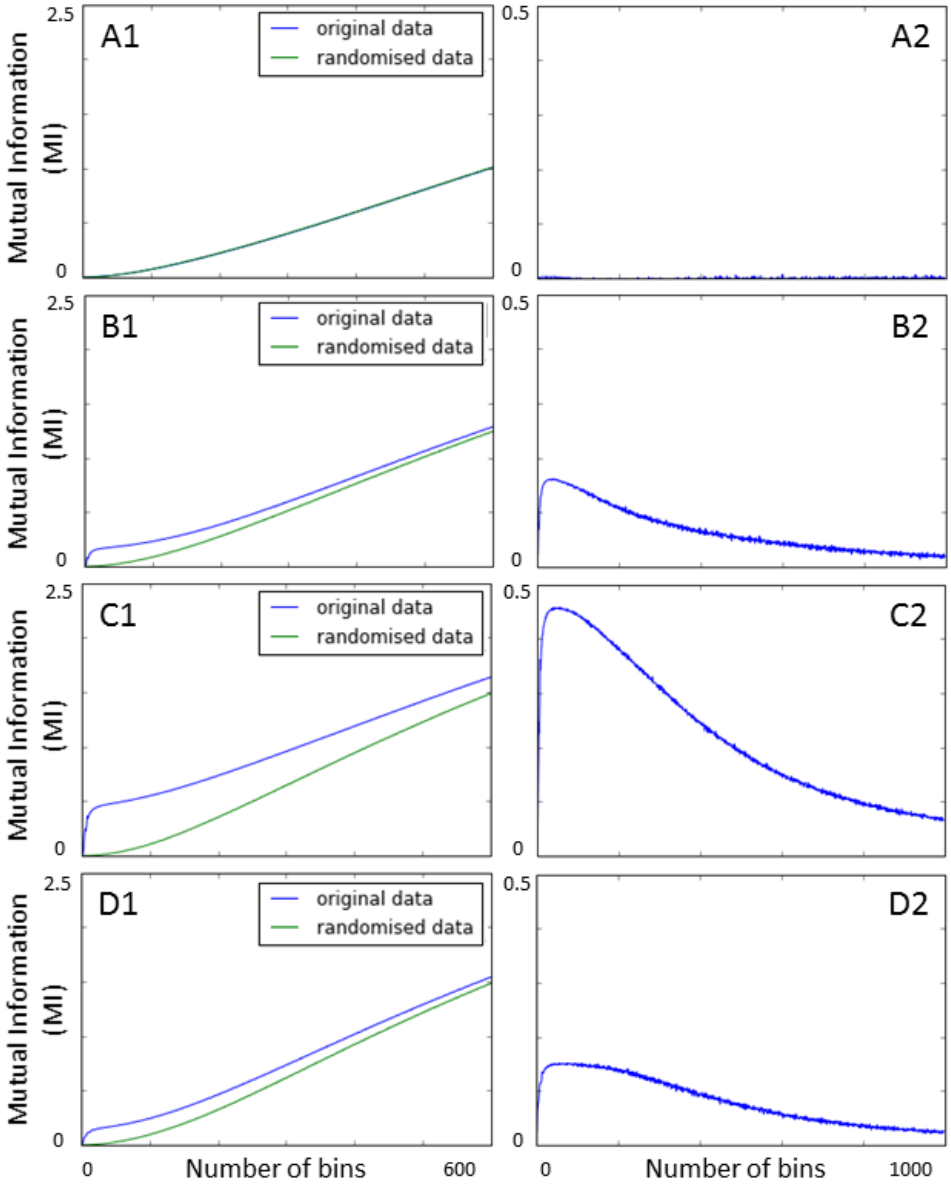


FIGURE 3.50: MI computed for a range of numbers of bins, for original data and with one set randomised in time in plot 1. Randomised MI subtracted from original data MI to give plot 2. A: Apo. B: inhibitor 1F8. C: activator 2A2. D: activator JS30.

	I(A;ATP(E_{int}))	***I(B;ATP(E_{int}))***	***I(C;ATP(E_{int}))***
APO	0.003	0.002	0.001
1F8	0.172	0.094	0.035
2A2	0.462	0.387	0.011
JS30	0.125	0.099	0.038

TABLE 3.11: MI computed between PC1 and distances A, B and C for four systems. Values reported as MI_{corr}.

Ligand	Scaffold	Activity	MI (protein)	MI (Pep-Thr)	MI (Prot+Pep)	MI (AllPep)
JS10	A	210	0.1499	0.2672	0.1403	0.1073
JS18	A	370	0.2697	0.2437	0.0680	0.2874
JS19	A	510	0.0494	0.0393	0.0327	0.0313
JS09	B	240	0.0505	0.0551	0.0099	0.0386
JS04	B	330	0.0490	0.0360	0.0663	0.0550
JS23	B	460	0.0797	0.0966	0.0320	0.1366
JS30	B	630	0.1458	0.0894	0.1300	0.1387
1F8	n/a	50	0.1464	0.0475	0.1489	0.0825
2A2	A	394	0.4740	0.2606	0.3284	0.1456
Apo	n/a	100	0.0073	0.0166	0.0159	0.0141

TABLE 3.12: MI_{corr} computed between various pairs of descriptors. For $MI = I(A;B)$, in all cases variable A is PC1. Variable B is the interaction energy of ATP with various parts of the system: protein, peptide, protein and peptide together, or peptide Thr only. Activity shown is as a percentage relative to Apo (where Apo is 100%).

3.3.7 Swapped structure trajectories

In order to validate that results are not caused only as an artefact of the crystal structure, two long simulations were run using crystal structures 3ORX and 3OTU, however swapping the ligands. Modelled protein from structure 3ORX was aligned to ligand JS30, and modelled protein 3OTU was aligned to ligand 1F8.

3.3.7.1 ATP γ-phosphate to Peptide-Thr distance

From observation of the resulting trajectories, after around 200 ns, the peptide dissociates from the active site for the inhibitor bound simulation (figure 3.51A). This is in line with results obtained from the simulation run for the inhibitor in the inhibited conformation; the peptide does not remain close enough to the active site for a period of time which would be required for the phosphate transfer to occur. For the activator bound simulation (figure 3.51B), there is some fluctuation of the peptide in the first 100 ns, which from observation of the trajectory is from the Thr end of the peptide flipping away from the ATP site, but remaining bound to the active site by the other end of the peptide. After this, the distances remain reasonably short, and large sections of the trajectory are stable at distances below 4 Å.

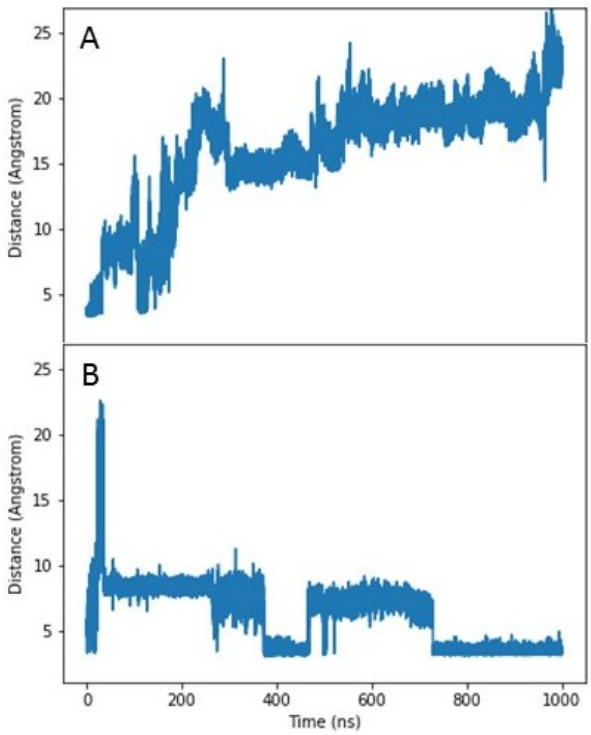


FIGURE 3.51: Substrate peptide Thr(O) to ATP (P-γ-phosphate) distance for swapped structure trajectories for A: inhibitor 1F8 bound to active conformation (structure 3OTU) and B: activator JS30 bound to inhibited conformation (structure 3ORX).

3.3.7.2 Specific distances: ATP γ-phosphate to Tyr54

From the original set of compounds (table 3.2A), the conformation of Tyr54 varied between the activator bound and inhibitor bound structures, with the distance between Tyr54 and ATP being larger in the inhibitor bound structure. From the swapped trajectories, it can be seen that in the case of the inhibitor bound simulation (figure 3.52A), the distance between Tyr54 and ATP begins at the distance of the activator bound structure, however after around 400 ns, the conformation flips, and resembles that of the inhibited structure. However for the simulation of the activator bound (figure 3.52B) to the inhibited structure this flip is not seen. It may be the case that repulsion of Tyr54 in the inhibited structure with residues around the active site are enough to cause this flip within the timescale of the simulation, but not the case for the activator bound simulation. The flipped out conformation seen in the inhibited structure is not as sterically crowded as the flipped in inhibited conformation (as seen in figure 3.29).

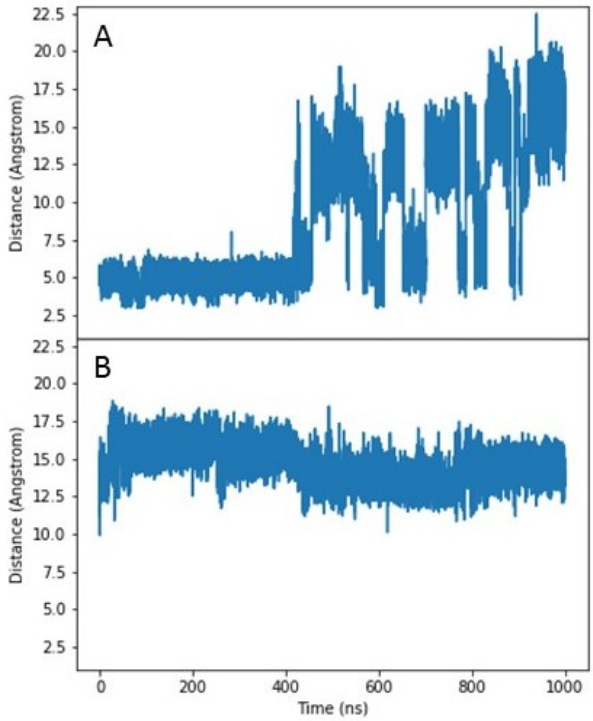


FIGURE 3.52: Tyr54(O) to ATP (P-γ-phosphate) distance for swapped structure trajectories for A: inhibitor 1F8 bound to active conformation (structure 3OTU) and B: activator JS30 bound to inhibited conformation (structure 3ORX).

3.3.7.3 Comparison to PCA on original compound set

Furthermore, the activation loop conformation obtained from the PCA in the high activating compounds is obtained by the simulation with the inhibited starting structure but with JS30 bound. This is shown in figure 3.53B, where low values of distance correspond to the activation loop close to helix C, which is a conformation of PC1 which was only seen in the higher activating compounds. For the simulation starting from the active conformation with inhibitor 1F8 bound, this conformation is not seen, as shown in figure 3.53A, where distances remain above 15 Å throughout the simulation.

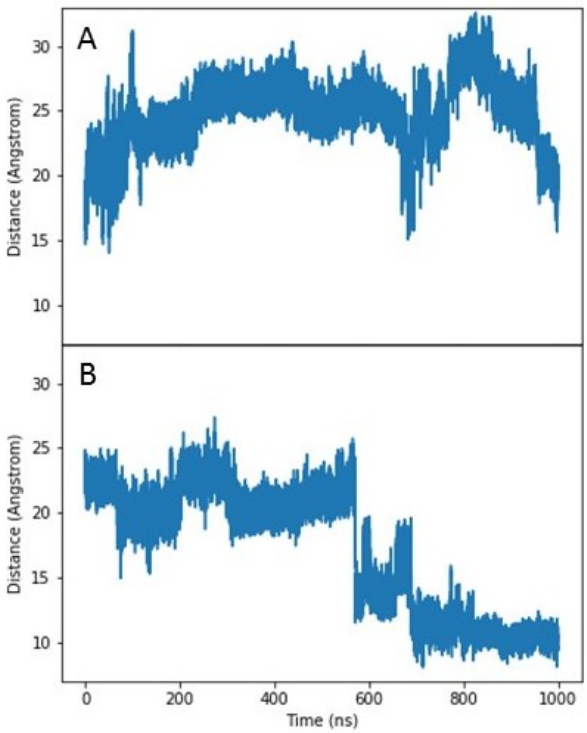


FIGURE 3.53: Helix C (Pro53) to activation loop (Lys163) distance for swapped structure trajectories for A: inhibitor bound to active conformation and B: activator bound to inhibited conformation.

The closing of the loop towards the helix C which is seen in the high activating compounds, seems to also be correlated to the position of the peptide, as the loop closes, the peptide-ATP distance is reduced. This can be quantified by considering the MI of these 2 distances, and results can be seen in figure 3.54.

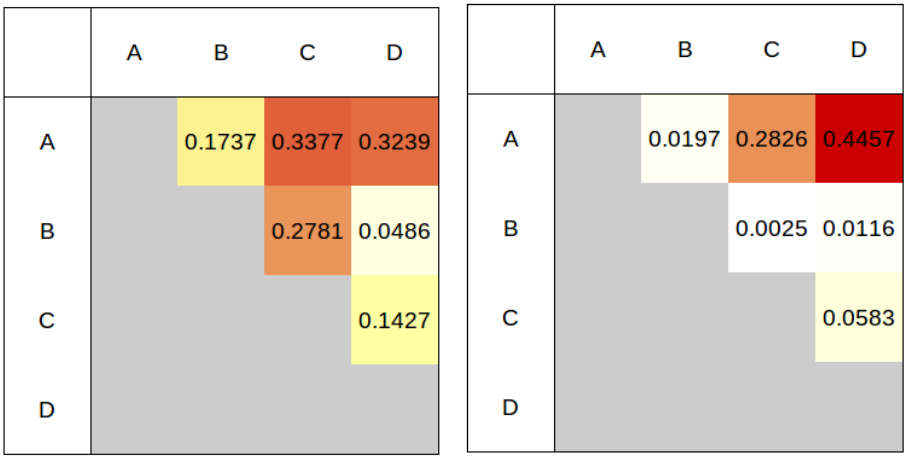


FIGURE 3.54: MI calculated between two distances for swapped simulations (A: inhibitor bound to act structure and B: activator bound to inh structure). Distances A: Peptide Thr to γ-phosphate of ATP, B: Lys39 to Glu58, C: Tyr(O) to γ-phosphate of ATP, and D: Activation loop to helix α-C. Higher MI is seen for $I(A;D)$.

3.4 Discussion

Initial analysis of PDK1 focused on particular distances which were selected based on known information from the literature. This highlighted that for inhibitor bound PDK1, the substrate Thr residue does not remain close enough to ATP to allow for phosphate transfer, however with any of the allosteric activators this distance was stable at shorter distances. For the inhibitor and activator 2A2 this was validated by repeat 100 ns simulations, and in all cases the substrate moves away from the active site with ligand 1F8 bound. The other distances computed showed potential differences when considering only the original set of four simulations, however when extended to the full set of compounds it does not seem possible to rank compounds depending on their level of activation.

The PCA and KL divergence analysis highlight the importance of the activation loop in the function of PDK1. The differences in the conformation of the activation loop were confirmed using both the KL divergence of dihedral angles, and with Cα coordinate PCA, which both show differences between activated and inhibited PDK1. When extended to the full set of compounds it was again difficult to establish a trend relating to different levels of activation. By then computing the interaction energies of ATP with the protein, it suggests that there is correlation between the loop motion and the interactions with ATP at the active site as values of MI between these two variables are higher than for simulations with no allosteric ligand bound. Further studies to confirm this hypothesis could be done by carrying out mutations of key residues which interact with ATP and the activation loop; such as Lys39, Glu58, or Tyr54.

To further understand the allosteric mechanism of PDK1 activation and inhibition, the work by Schulze et al. [88] should be investigated further. An overall description of the allosteric mechanism of PDK1 should include both the findings from this thesis; but also account for the changes in the global 'hinge-twist' motion, and the changes in the length of α-helix B, which were confirmed using enhanced sampling techniques in the work by Schulze.

Chapter 4

Small molecule allosteric effects on the WPD loop of protein tyrosine phosphatase 1B (PTP1B)

4.1 Introduction

4.1.1 Protein phosphatases as a drug target

Protein phosphatases carry out the reverse function to protein kinases, in that they dephosphorylate their substrates by hydrolysis of a phosphoester at Ser, Thr or Tyr residues. There are far fewer phosphatases than kinases: 200 phosphatases compared to around 518 kinases. Initially it was believed that phosphatases were not as useful drug targets as kinases to regulate aberrant phosphorylation controlled signalling pathways [126–128], as their regulatory function was not clearly understood, and so were overlooked while focus was on protein kinases. However both kinases and phosphatases are just the "on" and "off" switches in these signalling pathways: either can dysfunction, and either may require their activity to be modulated. Similarly to kinases, dysregulation of phosphatases has been implicated in many different disease pathways, including cancer, diabetes and obesity, and also for immune and neurodegenerative disorders [127, 129–131].

Phosphatases can either act on Ser/Thr, or Tyr, and some have dual specificity. Despite there being over 400 Ser/Thr kinases, there are remarkably few Ser/Thr phosphatases (PSPs), only around 30. These are split into three families: PPPs (phospho-protein phosphatases); PPMs (metal-dependant protein phosphatases) and a class of Asp based phosphatases. In the case of PPPs, the fact that so few phosphatases can regulate the activity of many different substrates, has been linked to their ability to bind many different regulatory domains, and in fact all within this class are multimeric proteins. PPMs in contrast, bind either Mg^{2+} or Mn^{2+}, however they contain domains other than the catalytic domain which may facilitate substrate selection [132, 133].

Protein tyrosine phosphatases (PTPs) account for the largest number of protein phosphatases. There are three main classes of PTPs, which are defined based on their catalytic residue. The majority (116 out of 125) have a cysteine as the residue which receives the phosphate from the substrate, and the remaining 9 have either aspartic acid, or histidine. Further classification of the cysteine phosphatases is based on sequence similarity and function. Class I includes around 95% of Cys-PTPs [128, 129, 134], and all have a conserved active site sequence. Class I is then further split into two subgroups: those which only dephosphorylate tyrosine, and those with dual specificity for Ser/Thr and Tyr. The remaining two classes (II-III) include only a few members.

Protein tyrosine phosphatases are implicated in a large range of disease pathways [135], depending on the particular PTP. For example, many are associated with cancer, such as CDKN3 [136], PTEN [137], DUSP16 [138] or CD45/PTPRC [139, 140]. Others such as PTP1B [141, 142] or LMPTP [143] are targets to treat diabetes or obesity . Neurodegenerative diseases such as Alzheimer's and Parkinson's also may be linked to function of PTPs such as STEP [144] or SYNJI [145].

Yet while there has been significant success in development of kinase inhibitors (over 30 with FDA approval), there have been no phosphatase inhibitors approved, and only a very small number have reached clinical

trials. The issues are mostly with selectivity, since the active site is well conserved across many phosphatases in both sequence and structure and many considered PTPs to be "undruggable" [146, 147]. In addition to selectivity, P-Tyr mimetics which bind to the active site require them to be charged molecules, as the side chains which bind P-Tyr are positively charged. This then also leads to issues with cell permeability [130, 148].

Many allosteric sites on PTPs are known, and often these sites are less conserved between close family members. In addition, there is then the opportunity to select sites which do not require such charged molecules.

4.1.2 PTP1B Protein tyrosine phosphatase 1B

PTP1B is a tyrosine phoshpatase; the initial member of the class I PTPs and the first to be confirmed as a therapeutic target [149]. It is encoded by the PTPN1 gene, and as it negatively regulates the insulin pathway, it has been of interest as a target for treatment of obesity and diabetes. In studies of PTP1B-knockout mice, this role was confirmed, and mice show resistance to both obesity and diabetes [150]. However there is also interest in targeting PTP1B for a range of other conditions, including liver diseases [151] and cancer [152–154].

Initial developments of active site inhibitors yielded several compounds which have structures based on the P-Tyr substrate of PTP1B, or those which bind to both the active site and an adjacent binding pocket [130]. However it still remains that the lack of selectivity by using P-Tyr mimetics is a major challenge, and as a result, focus has shifted to allosteric sites. However as of yet only a very small number (3-4) of PTP1B inhibitors have reached the clinic. Those which have, were stopped in majority due to selectivity issues, such as Ertiprotafib, which reached Phase II trials, but off target binding resulted in toxicity issues [155]. Also Trodusquemine showed some promise, trials were previously suspended again due to selectivity concerns. However it is possible that research on this compound will recommence as Novo Biosciences were awarded further funding to continue investigating the scope for this compound [128, 155, 156].

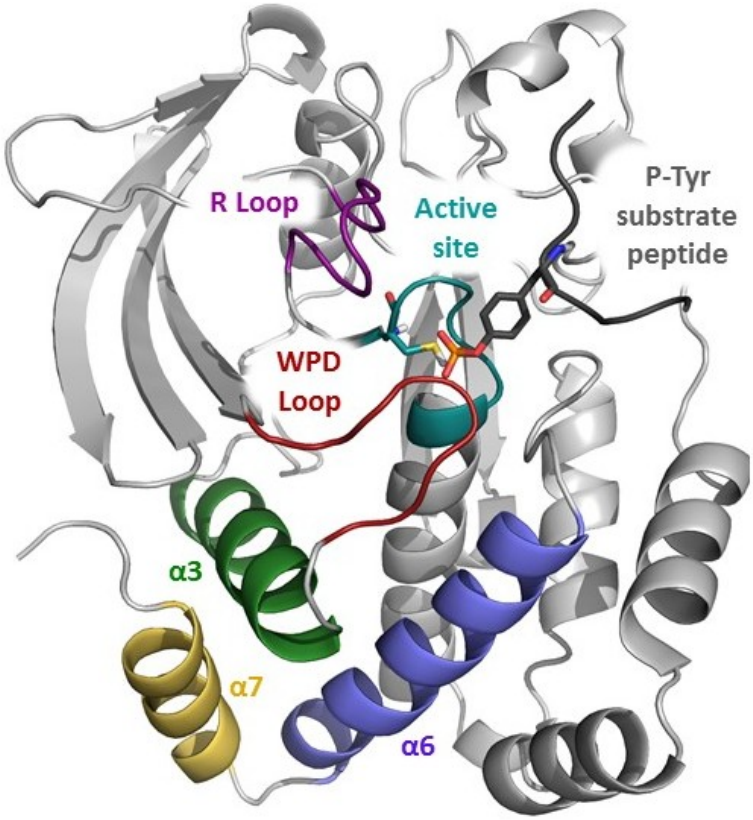


FIGURE 4.1: Structural features of PTP1B. Phophate is transferred from the P-Tyr substrate to Cys215 at the active site (teal). The WPD loop must be open for the substrate to bind, and closes over the substrate to position key residues for catalysis.

The structure of PTP1B is shown in figure 4.1, highlighting the important structural features. The active site contains an arginine residue, which facilitates substrate P-Tyr binding and allows for nucleophilic attack by Cys215, illustrated in figure 4.2. In order for the substrate to access the active site, the WPD loop moves to an open conformation, which allows the substrate to bind, after which the loop closes (figure 4.3), and interactions of an aspartic acid on the WPD loop with the substrate allow for phosphate transfer.

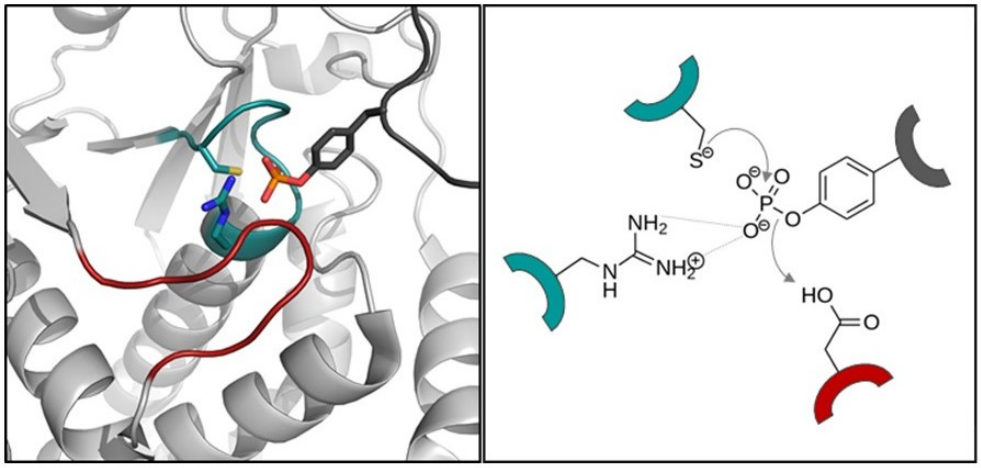


FIGURE 4.2: Mechanism of dephosphorylation of substrate. Arg221 facilitates substrate binding. Phosphate is transferred to Cys215. Asp181 on the WPD loop provides H^+ in substrate-phosphate bond breaking.

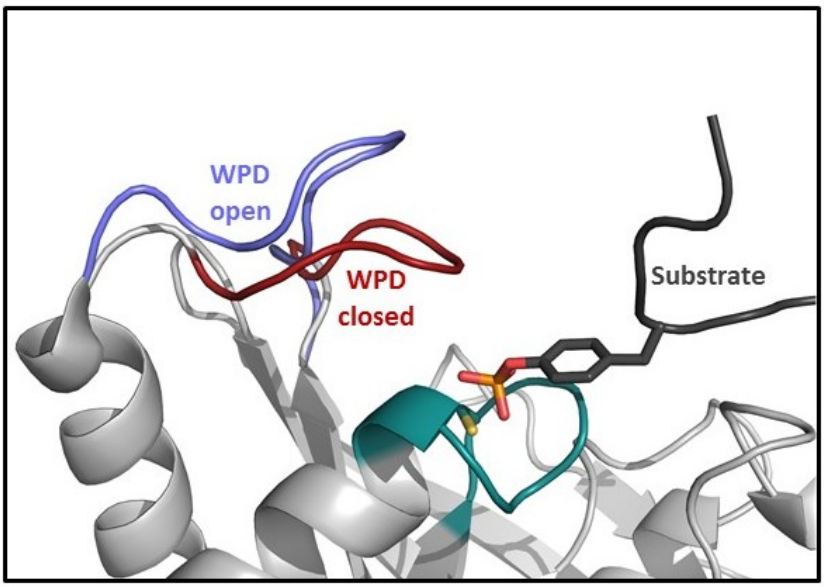


FIGURE 4.3: Open and closed conformations of the WPD loop.

PTP1B has two currently known allosteric sites, one which is a pocket formed by the α-3, α-6 and α-7 helices [157]. The other site was more recently discovered [158], and sits on the other side of α-3 as shown in figure 4.4. The current understanding is that binding of an inhibitor stabilises the inactive conformation of the WPD loop (figure 4.1), however it is unclear exactly how this stabilisation occurs. Some mechanistic understanding can be obtained for specific ligands, however there currently seems to be no way to apply this to rational design of new compounds.

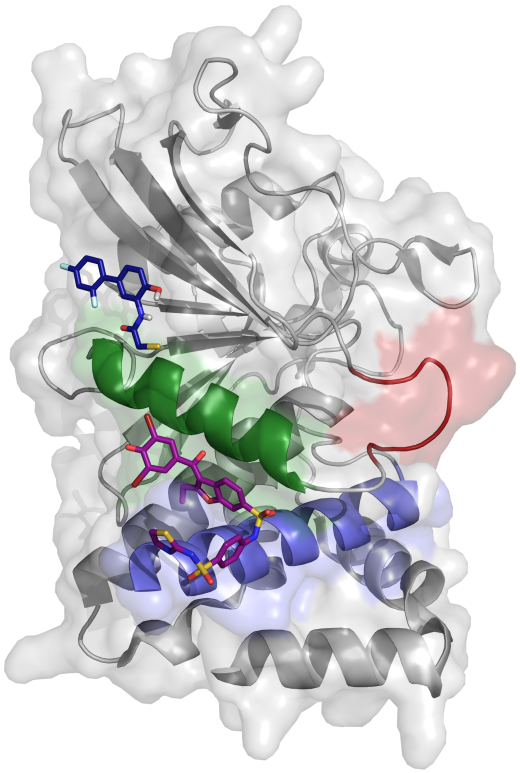


FIGURE 4.4: Two known allosteric sites of PTP1B. Inhibitor FRJ (PDB ID 1T4J) shown in purple, and inhibitor D0P shown in blue (PDB ID 6B95). Colours of structural features of PTP1B as in figure 4.1

4.2 Methods

4.2.1 Molecular modelling

4.2.1.1 Ligands

Simulations were set up for two different inhibitors, which are bound to two different allosteric sites. The first is from crystal structure PDB ID 1T4J and the second from PDB ID 6B95 (figure 4.4).

For 1T4J, partial charges were derived using the AM1-BCC methodology using Antechamber [98]. For 6B95, the ligand is covalently bound to the allosteric site, and methods described in section 3.2.1.1 were used to describe the ligand-cysteine bond.

4.2.1.2 Protein preparation

Crystal structures were found in the PDB database, PDB ID 1SUG (closed WPD loop), 2HNP (open WPD loop), 1T4J (open loop with inhibitor FRJ) and 6B95 (open loop with inhibitor D0P). The sequence used for all models includes residues Met1 to Leu299. All His were modelled as HIE except His214 which is modelled as HID, as this residue forms an H bond network with Tyr124 and His173. For structure 1SUG, only residue Met1 was missing. For structure 2HNP, α helix 7 was missing (residues 283-298) and so coordinates for these residues were taken from 1SUG. This was validated by modelling these using the multi-structure alignment of MODELLER [100] and the helix is an identical position. For structure 1T4J, residues 284-289 were missing, and were modelled using MODELLER. As residues 290-298 are present in the crystal structure, it is clear that with inhibitor bound, α helix 7 is disordered, as the ordered helix of 1SUG would conflict with the binding site of the allosteric inhibitor. For structure 6B95, residues 279-299 were modelled using MODELLER. This structure has three mutations relative to the other structures of PTP1B. One of these is necessary to retain, as it is the Cys residue which the allosteric ligand is covalently bound to. The other two mutations have been altered in order to compare directly with the other simulations. In all cases, crystal waters and ions were removed, and

the protein structure was prepared using Maestro [101]: missing hydrogen atoms were added, and N-methyl and acetyl groups were added to the C and N terminal ends of the protein respectively.

4.2.1.3 Substrate peptide

Three peptide substrate complexes have been found for PTP1B, and have available crystal structures (1EEN, 1EEO, 4ZRT). The peptide with sequence ACE-ELEF-(Y-phos)-MDYE-NH_2 was selected for simulations, using coordinates for the peptide from crystal structure PDB ID 1EEO. In all cases, the substrate to active site distance was kept reasonably constant at the start of the simulation (between 2.2 - 2.8 Å).

4.2.1.4 Ligand substrate protein complexes

In each case, a complex was set up including protein along with the substrate peptide or substrate peptide with allosteric ligand using the software leap from the Amber16 software package [159]. General Amber Force Field parameters were assigned to ligand atoms with the addition of the adapted disulphide bond parameters discussed previously, while the FF14SB-ILDN force field [45] was used to describe the protein. Phosphate parameters developed by Case et al. [112] were used to describe the phosphotyrosine located on the substrate peptide. Each model complex was then solvated in a box of TIP3P water molecules extending 10 Å from the edge of the solute, and Na^+ ions added to neutralise the net charge of the complex.

General Amber Force Field parameters were assigned to ligand atoms with the addition of the adapted disulphide bond parameters (as per method described in section 3.2.1.1) for the simulation based on structure PDBID 6B95, while the FF14SB-ILDN force field [45] was used to describe the protein. Phosphate parameters developed by Case et al. [112] were used to describe the phosphotyrosine of the substrate peptide.

4.2.2 Molecular dynamics simulations

4.2.2.1 Equilibrium MD simulations

The resulting solvated models were energy minimised with sander, and equilibrated at NVT using PMEMD (CUDA), from the software package Amber16 [159]. Energy minimisation using 200 steps of conjugate gradient with restraints on the solute was carried out in order to equilibrate only the solvent. This was then followed by 5500 steps of conjugate gradient followed by 1500 steps of steepest descent, with restraints only on the WPD loop and the substrate. The system was then heated to 298 K in four steps using the Berendsen thermostat [52], each of 100 ps, with harmonic Cartesian positional restraints of 10 kcal mol^{-1} $Å^{-1}$ on the substrate and on the WPD loop. A further 5 ns was run in the NPT ensemble while retaining the restraints, in order to equilibrate the density of the system. The restraints were then removed over a further 800 ps at NVT. A further 5 ns in the NPT enemble was run prior to the production run. Equilibrium molecular dynamics simulations were run using PMEMD (CUDA), for a simulation time of 1 μs using a 2 fs timestep. Throughout all equilibration and production MD, SHAKE was applied to constrain all bonds involving hydrogen. PME was used to describe the long range electrostatic interactions using a cutoff of 8 Å. In all simulations, snapshots were saved every 5 ps, resulting in 200k snapshots for every 1 μs simulation.

Simulations were started from both open and closed WPD loop in each case. Systems set up were: substrate only, substrate with inhibitor FRJ (1T4J) and substrate with inhibitor D0P (6B95). Simulations without substrate or allosteric ligand were also run for comparison.

4.2.2.2 Steered MD simulations

In order to obtain intermediate conformations between the open and closed WPD loop states, steered MD (sMD) was performed using Gromacs [160] with Plumed [161]. Starting structures for the substrate bound open and closed conformations were taken from the PDB, using structures PDB ID

2HNP for the open conformation and PDB ID 1SUG for the closed conformation. The conformation of Trp179 was altered in structure 2HNP (see section 4.3.1) by adjusting both chi1 and chi2 torsional angles. The substrate used was the same as described in section 4.2.1.3.

4.2.2.3 Seeded equilibrium MD simulations

Coordinates were extracted from the steered MD simulations as starting points for many short equilibrium MD simulations. RMSD of the WPD loop was monitored for the sMD simulation and 200 structures representing a range of values of RMSD were selected. For each structure, minimisation, equilibration and production MD simulations were run using the same conditions detailed in section 4.2.2.1.

4.2.3 MSM generation

A Markov State Model (MSM) was constructed using both the equilibrium MD simulations and seeded simulations from the steered MD, using the pyEMMA software package (version 2.5.4) [116]. In order to reduce the dimensionality of the dataset and define separate states, descriptors were selected which highlight differences between the active and inactive conformations. The RMSD of the WPD loop (residues 180-185 were included) relative to the their position on the open WPD loop structure, and the substrate to Cys215 distance were selected as input dimensions. K-means clustering using 100 clusters was used to define a set of microstates. In order to obtain a separation of timescales for both the substrate and substrate plus inhibitor data sets, initial clustering was completed separately on each set of data. A range of implied timescales were computed using lag times between 1 and 10000 steps. A lag time of 3000 steps (30 ns) was selected for the initial model, based on plots generated for a range of implicit timescales, as shown in figure 4.22. Three macrostates were then selected to define the slow processes. The k-means clusters were separated into three groups with a hidden Bayesian Markov Model [162].

A second model was constructed, again using a k-means clustering of 100 clusters, however clustering was performed on the combined set of simulations including the substrate only simulations, and the substrate with inhibitor simulations. A lag time of 2000 steps (20 ns) was selected. For this model, assignment of the 100 microstates into three macrostates was done using PCCA [163, 164].

4.3 Results

4.3.1 Protein structures

Initial models based on the structure of PDB ID 2HNP highlighted potential inaccuracies with the conformation of the side chain of Trp179 when carrying out the loop closure using sMD. By applying only a bias of the WPD loop RMSD to the closed conformation, it was not possible to close the loop. This was due to the conformation of Trp179 which must be flipped before the loop closed, as it is flipped in the closed structure, and obstructs the loop closure (figure 4.5). If the loop is even partially closed, there is insufficient space for this flip to occur. Further attempts at sMD included bias based on both the loop RMSD and the chi1 and chi2 torsions of Trp179 residue, and as long as the Trp had flipped conformation *before* closing the loop, this was possible. RMSD plots of this process are shown in section 4.3.2.

However in more recently released inhibitor bound crystal structures (6B95 and related structures), it was apparent that none had the initial Trp conformation of 2HNP. A further search of the PDB found no other PTP1B structures with the same conformation of this side chain. On checking the electron density of 2HNP it became apparent that this Trp conformation was not visible at all in the electron density. It seems likely that this is an error, since other structures which show better resolution have the flipped conformation. Therefore further sMD was carried out with the Trp in the flipped conformation shown in dark grey in figure 4.5, by manual adjustment of the model constructed from 2HNP, and only loop RMSD used as a bias for the steered MD. Several previous equilibrium MD and steered MD use 2HNP as a starting point, and as this conformation of Trp affects the loop closure, this could affect the outcome [165–168].

4.3.2 Steered MD simulations

With the unaltered structure of PDB ID 2HNP, initial attempts at steered MD to close the WPD loop resulted in reaching only a minimum RMSD of around 2 Å, or with an increased force constant caused the protein to unfold.

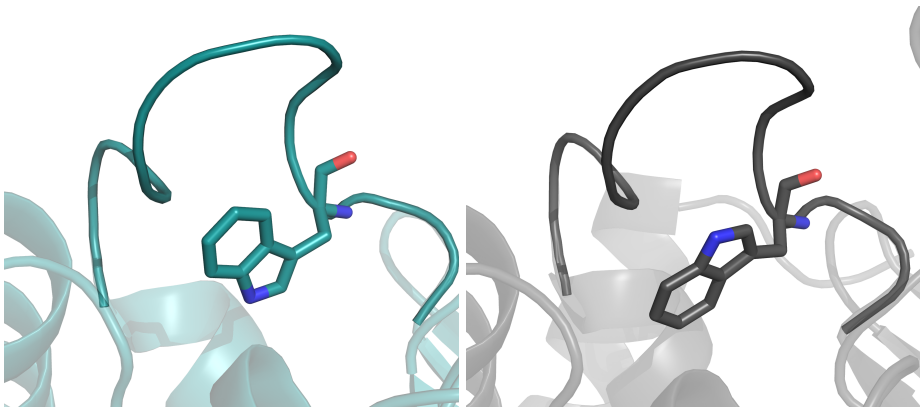


FIGURE 4.5: Conformation of Trp179 in PDB ID 2HNP (teal), and the proposed corrected conformation (grey).

Plots showing RMSD of the WPD loop over time can be seen in figure 4.6 for these attempts.

As was discussed above, the issue with the loop closure is the initial conformation of Trp179 in PDB ID 2HNP, which must flip prior to loop closing. As this caused problems when using the RMSD of the loop alone, additional steps were added to the steered MD to bias the motion of the $\chi 1$ and $\chi 2$ torsions of Trp179. Provided the flip occurred in the initial stages of WPD loop closure, it was possible to reach RMSD 1 Å over a 150 ns sMD simulation. However this later proved to be unnecessary, as it is very likely that the Trp conformation obtained from crystal structure 2HNP is incorrect. Bias using RMSD alone was then possible, to reach a final RMSD of around 1 Å.

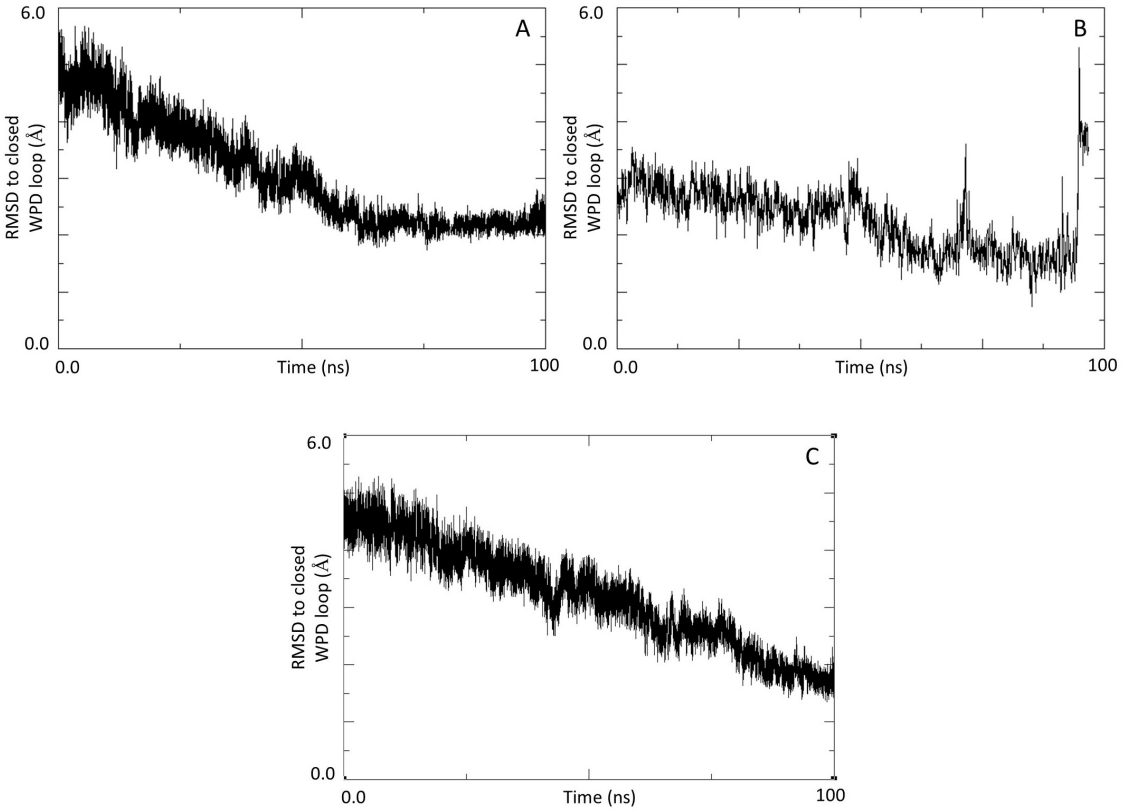


FIGURE 4.6: Steered MD simulations. A: unaltered structure of 2HNP using a force constant of 2500 kJ mol^{-1}. B: unaltered structure of 2HNP using a force constant of 3500 kJ mol^{-1}. C: altered Trp179 conformation of 2HNP structure using a force constant of 2500 kJ mol^{-1}.

4.3.3 Loop conformation

Long MD of four complexes were initially analysed. Codes in brackets shown below are used to refer to each of these simulations in the following discussion.

- Substrate bound: open WPD loop **(SO)**.
- Substrate bound: closed WPD loop **(SC)**.
- Substrate and inhibitor FRJ bound: open WPD loop **(IO)**.
- Substrate and inhibitor FRJ bound bound: closed WPD loop **(IC)**.

Of these four simulations from observation of the trajectory, only one visited conformations of the WPD loop which vary from its start point (i.e. open to closed, or closed to open). In the IC simulation, the loop opens at around 500-520 ns, as in figure 4.7, and remains open for the next 500ns.

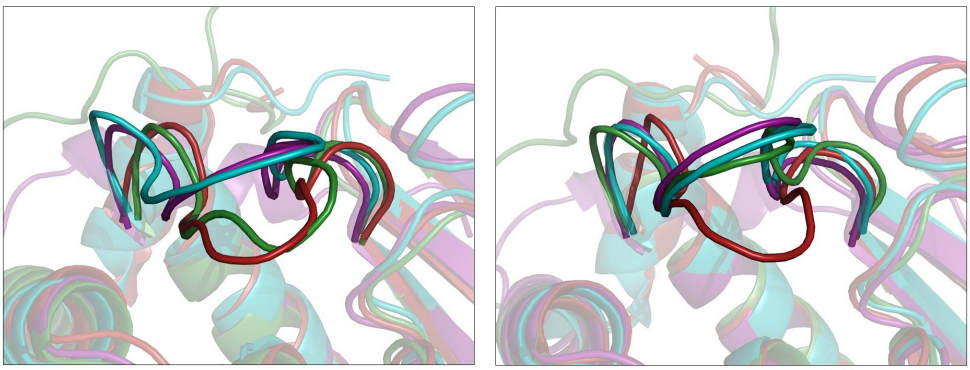


FIGURE 4.7: Snapshots of all four systems at 500 ns and 520 ns. Teal: SO. Red: SC. Purple: IO. Green: IC.

Plots showing the RMSD per snapshot, and distributions of RMSD for each simulation were then plotted, to highlight differences in behaviour.

In figure 4.8, comparison of the SO and the IO simulations show only a small shift in the distribution of values, and both show a reasonably large range of values suggesting the loop is reasonably flexible in both cases. The SC simulation shows a much more narrow range of values, and the loop is

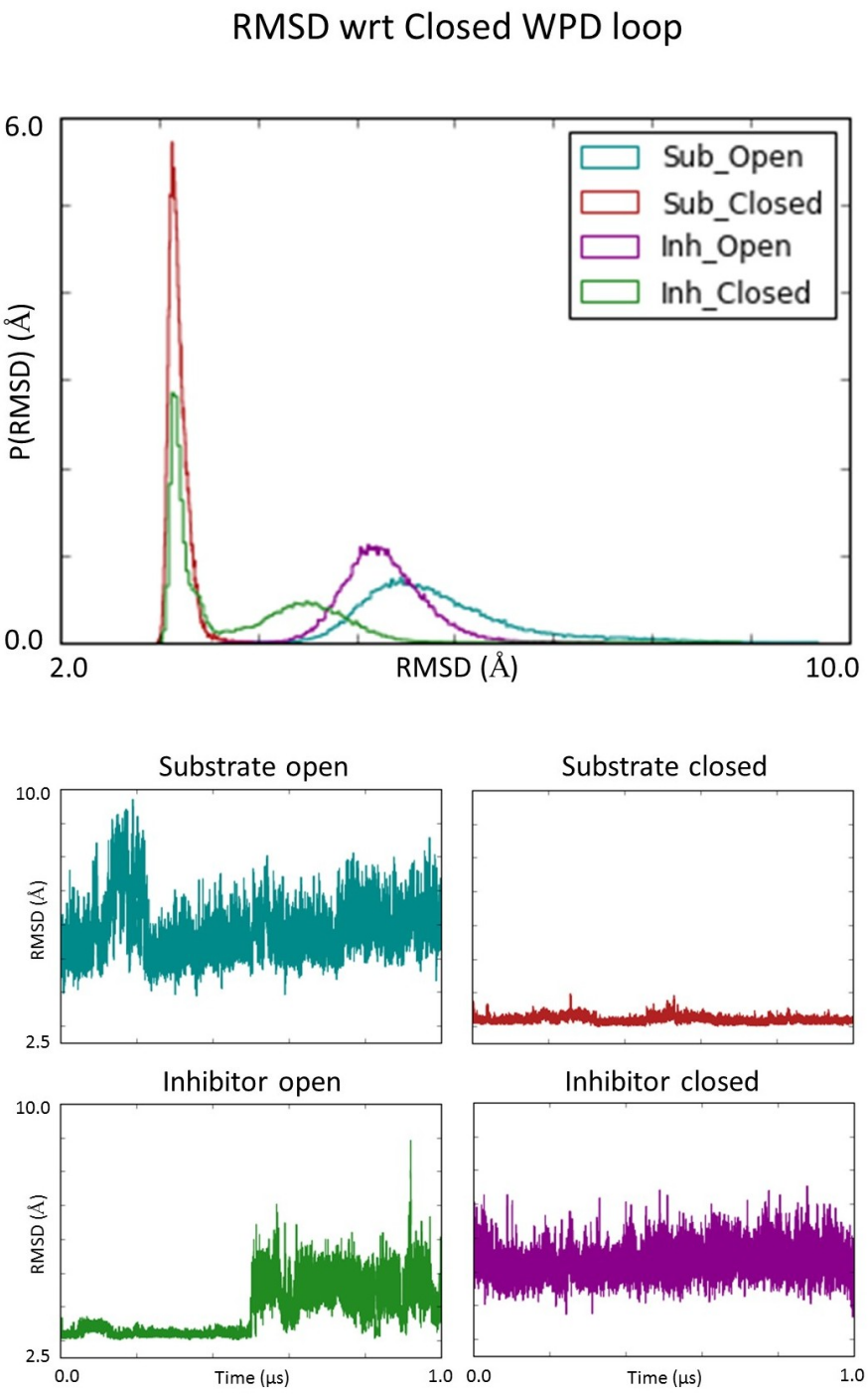


FIGURE 4.8: RMSD of residues Pro181-Pro186 relative to the closed loop conformation. Teal: SO. Red: SC. Purple: IO. Green: IC.

much more stable in the closed conformation, which can be seen from the time plot. However it can be seen in the time plot for the IC simulation, that at around 500 ns the RMSD relative to closed starts to change.

This is mirrored in the same plots, but with RMSD calculated relative to the open conformation, as seen in figure 4.9. This suggests that the difference in activity is due to the destabilisation of the closed state when inhibitor is bound. In both cases, the distributions of the open state seem unaffected by whether there is inhibitor bound or not.

For each set of RMSDs computed (relative to closed or open), the JS divergence was calculated for each pair of simulations, and results are in figure 4.10. This confirms that relative to the closed conformation, the SC simulation has a larger JS divergence to both closed conformations than the IC simulation. It also confirms that both the substrate and inhibitor bound open conformations are similar, as JS divergence values are very small. The same can be seen in the plot relative to the open conformation, and the JS divergence values of the inhibitor closed state take values between the maximum and minimum JS for any pair. This is because the inhibitor closed simulation has part of the distribution in the open state, and part in the closed state. Distributions of RMSD for residues Thr177-Glu186 were also computed and can be found in appendix C.

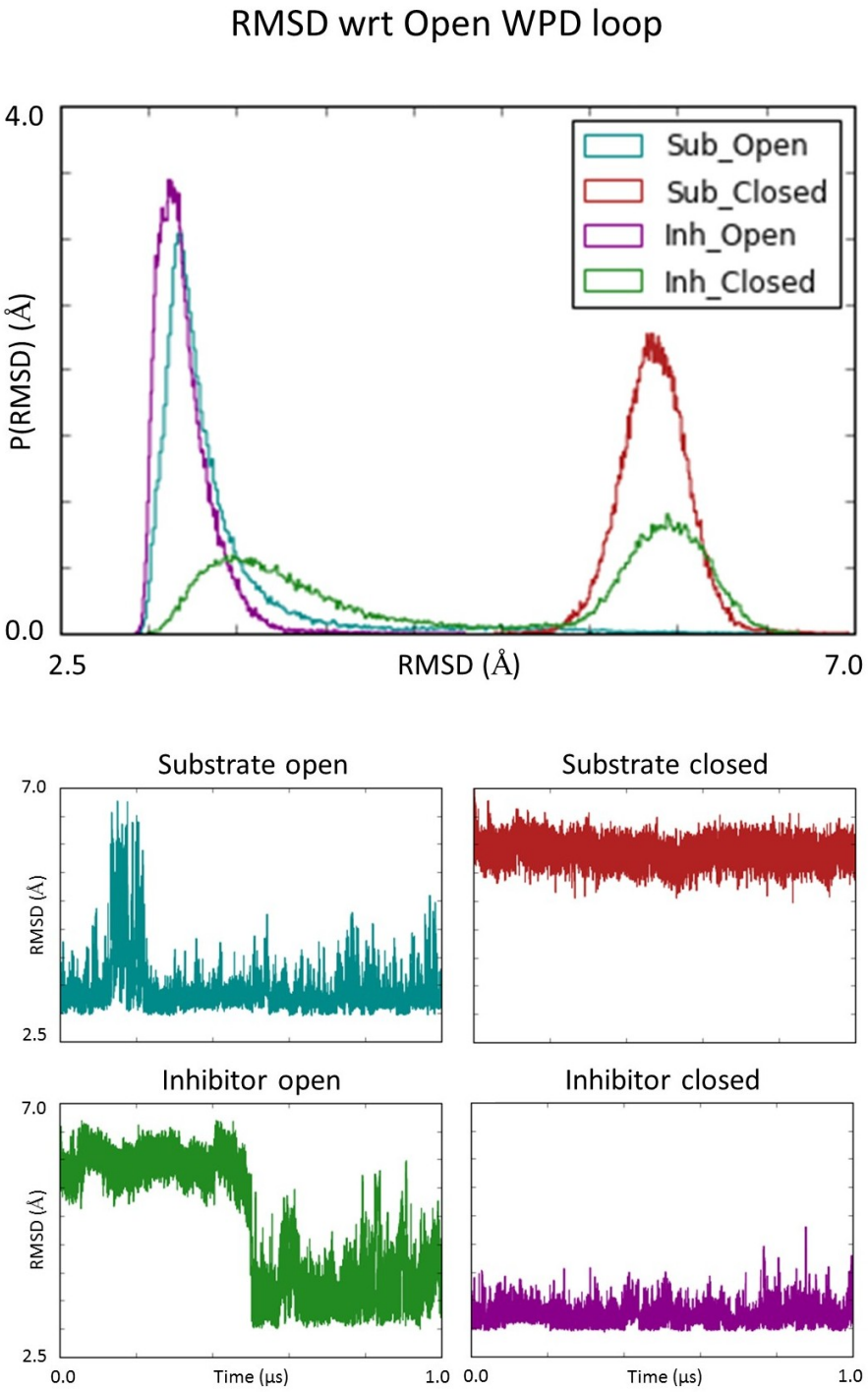


FIGURE 4.9: RMSD of residues Pro181-Pro186 relative to the open loop conformation. Teal: SO. Red: SC. Purple: IO. Green: IC.

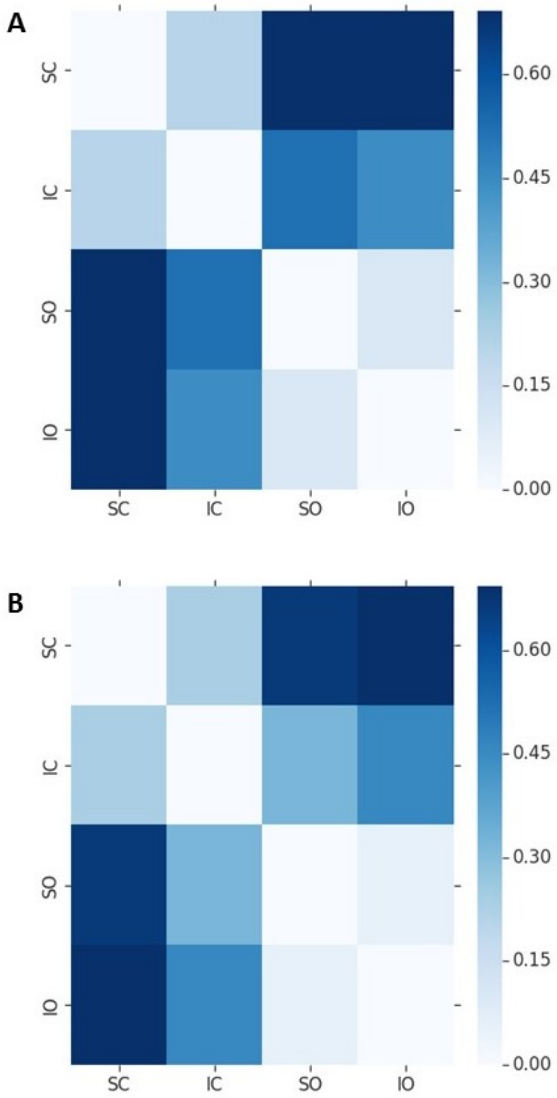


FIGURE 4.10: JS divergence of distributions of RMSD computed relative to A: closed and B: open. Clustering described in section 3.2.8, using ϵ=0.15 with 2 states.

4.3.3.1 Extending analysis to include inhibitor D0P

Further simulations were completed for allosteric inhibitor D0P, which binds to an alternate allosteric site to ligand FRJ. The RMSD analysis completed above was then repeated for these simulations. Simulations with both substrate and ligand D0P were started from both the closed and open WPD loop conformations, and RMSD relative to the closed and open conformations were computed.

As with ligand FRJ, the closed conformation with ligand D0P was not stable, and within around 100 ns the loop opens for this simulation. Figures 4.11 and 4.12 show distributions and time series for the simuations with ligand D0P bound. RMSD is computed for residues Pro181-Pro186. Distributions of the substrate open (SO) and substrate closed (SC) RMSD are also shown as a reference.

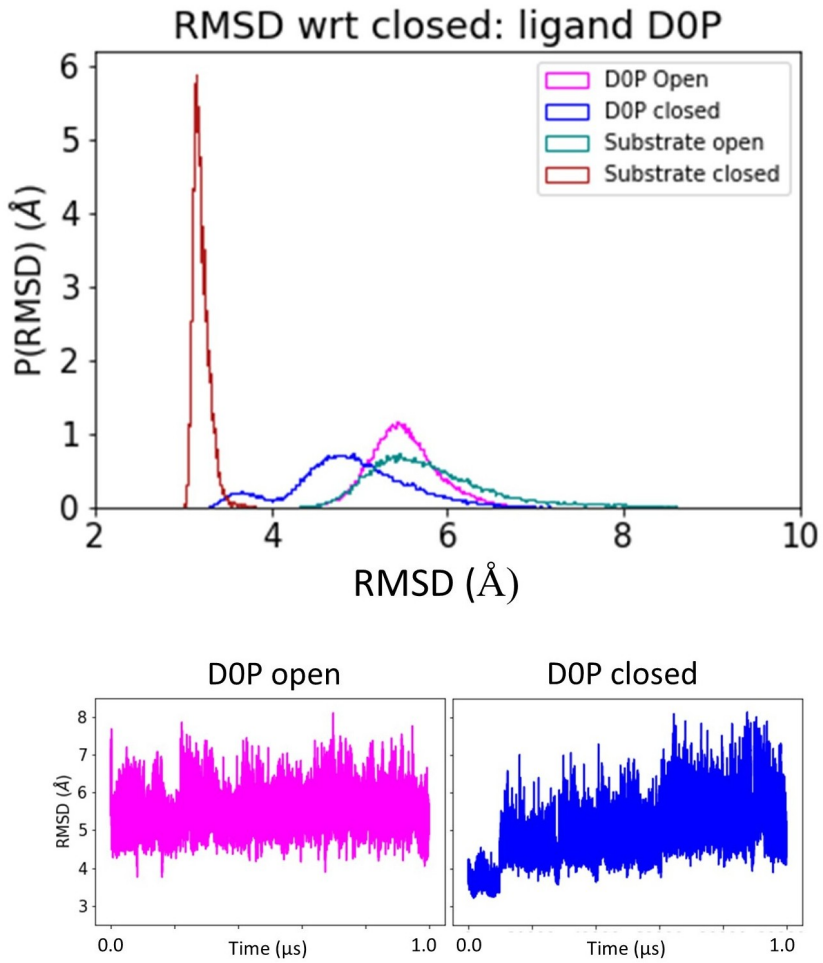


FIGURE 4.11: RMSD of residues Pro181-Pro186 relative to the closed loop conformation. Teal: SO. Red: SC. Magenta: D0P open. Blue: D0P closed.

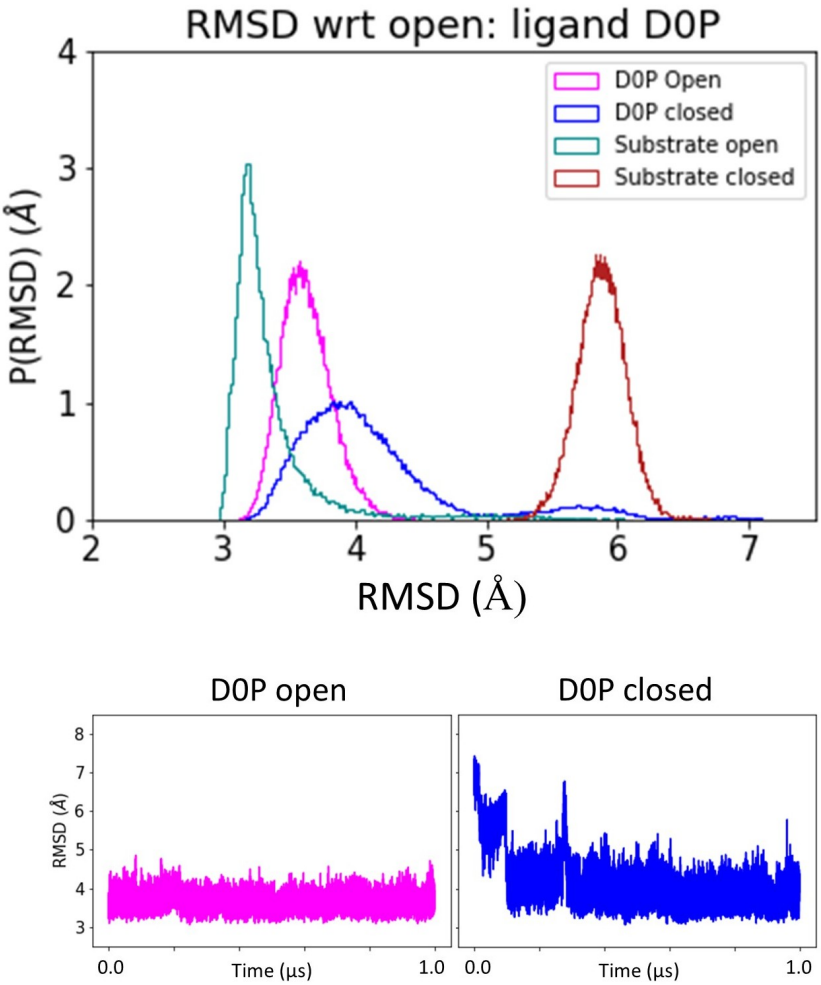


FIGURE 4.12: RMSD of residues Pro181-Pro186 relative to the open loop conformation. Teal: SO. Red: SC. Magenta: D0P open. Blue: D0P closed.

4.3.4 Equilibrium MD KL analysis

Equilibrium MD simulations for the SO, IO, SC and IC simulations were compared using KL divergence of torsional angles. Comparison of the SO and IO simulations shows that the largest differences in backbone torsions are around the active site, as shown in figure 4.13B1 and B2. The KL values for the sidechain torsions highlighted four residues located around the allosteric site and near to the WPD loop, namely Tyr153, Try154, Asn194 and Glu277. The difference in Glu277 is understandable, as this sits directly in the allosteric site.

For the comparison of the SC and IC simulations, the result is less clear, and residues with the largest backbone and sidechain KL values are located away from both the active and allosteric sites.

This emphasises the need for a multi-approach analysis, as while KL of torsional angles can give useful insights in some cases, it may not always be the case that large enough differences are seen.

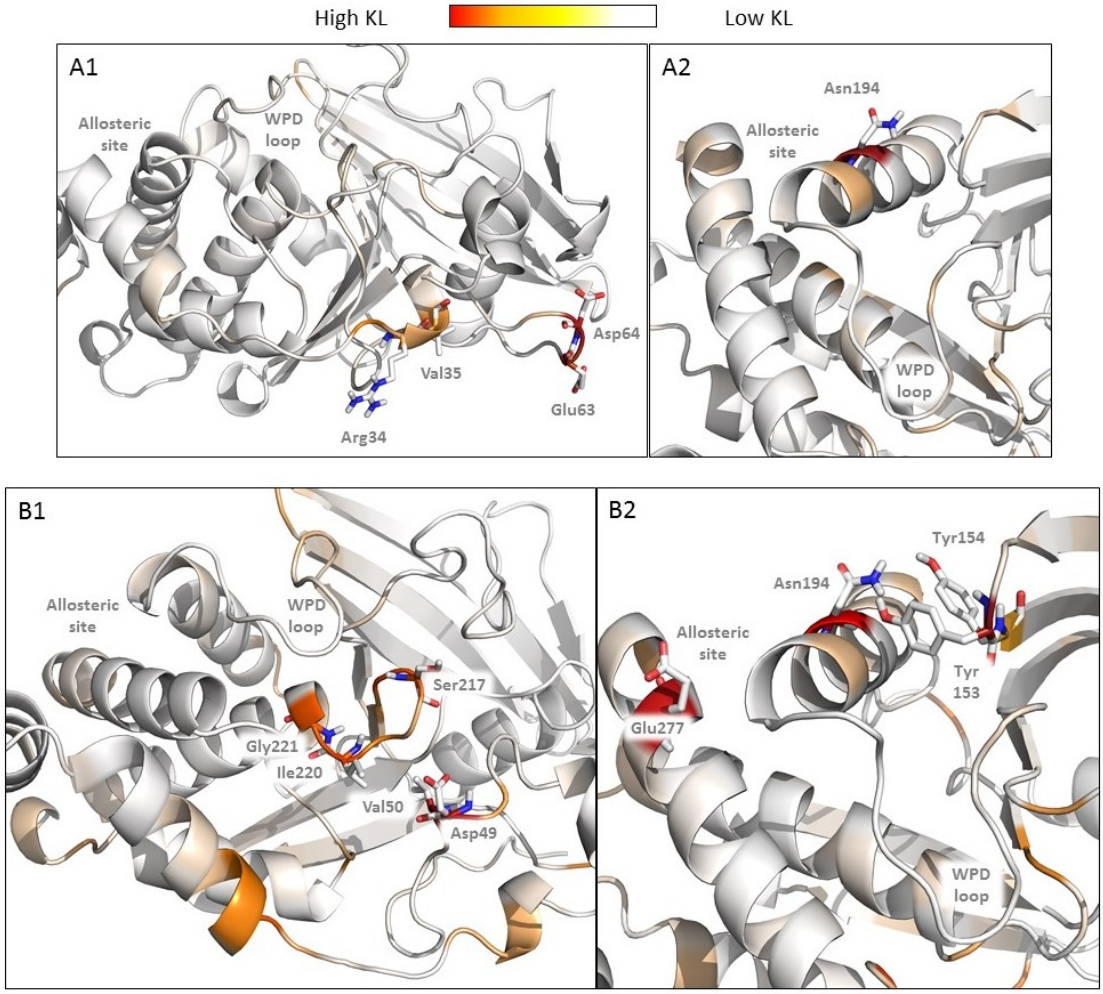


FIGURE 4.13: KL divergence computed for the open and closed conformation simulations. Figures show KL results for A1: $KL_{backbone}(SC||IC)$. A2: $KL_{sidechain}(SC||IC)$. B1: $KL_{backbone}(SO||IO)$. B2: $KL_{sidechain}(SO||IO)$.

4.3.5 Distributions of distances relating to reaction mechanism

Several distances were calculated based on known information about the mechanism. As highlighted in figure 4.2, the substrate P-Tyr must remain at a distance to Cys215 to allow phosphate transfer to occur. Furthermore, Asp181 on the WPD loop should interact with the substrate, as deprotonation of this residue is required during the breaking of the substrate-phosphate bond. Distributions of these distances were therefore calculated for each system to allow differences between substrate and inhibitor bound simulations to be assessed.

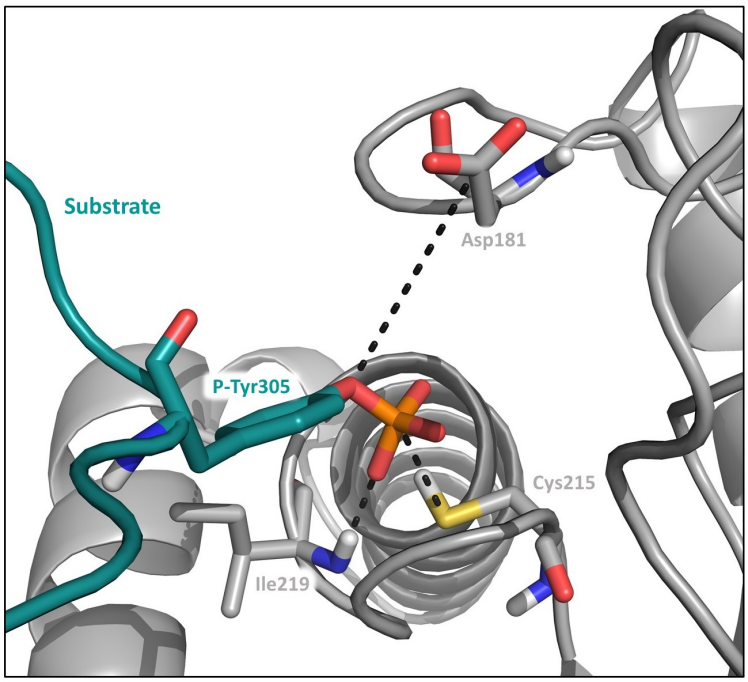


FIGURE 4.14: Three distances computed for each system. Asp181(C) to P-Tyr(O); Cys215(S) to P-Tyr(P); Ile219(N) to P-Tyr(P).

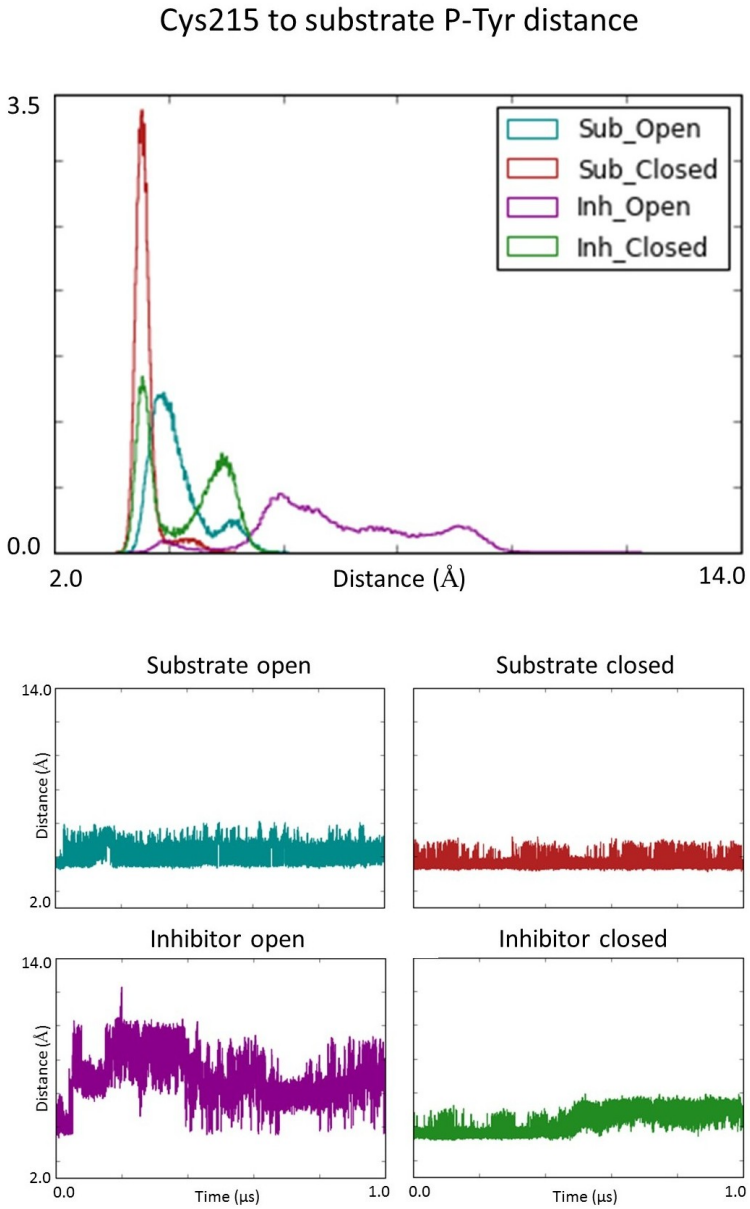


FIGURE 4.15: Distance from Cys215 to substrate P-Tyr. Teal: SO. Red: SC. Purple: IO. Green: IC.

The substrate remains at reasonably short distances (< 5 Å) in the SO, and SC simulations, with values for the SC simulation remaining in a narrow distribution between 3 and 4 Å. In the SO simulation, the distribution is broader, and maximum values increase to around 6 Å. With inhibitor bound, the closed conformation shows two peaks in the distribution of values, which correspond to the two segments of the trajectory: up to 500 ns before the loop opens, and after the loop begins to open. This can be seen from the time plot in figure 4.15, where at around 500 ns the value increases. With the IO simulation, the substrate only remains close to the active site for around 30 ns. After this time, the substrate remains bound to PTP1B, but moves noticeably away from the active site. The comparison of the SO and the IO highlights a potential destabilisation of binding of the substrate, when allosteric ligand is bound. The distance of Asp181 to the active site was measured as Asp181 is involved in the reaction mechanism, as detailed in figure 4.2. From the distributions shown in figure 4.16, it can be seen that the open conformation for the substrate and inhibitor vary, however the closed conformation shows a larger difference due to the loop opening in the IC simulation. The Ile219 to substrate distance (figure 4.17) shows a similar pattern to the Cys215 to substrate distance, as they both relate to how close the substrate P-Tyr is to the active site. However in this case, the distribution of the IC simulation is at shorter distances than the SC simulation, in contrast to the Cys215 distance. This could result in different positioning of the substrate in the active site, and not allow the shorter Cys215-P-Tyr distances required for catalysis to occur.

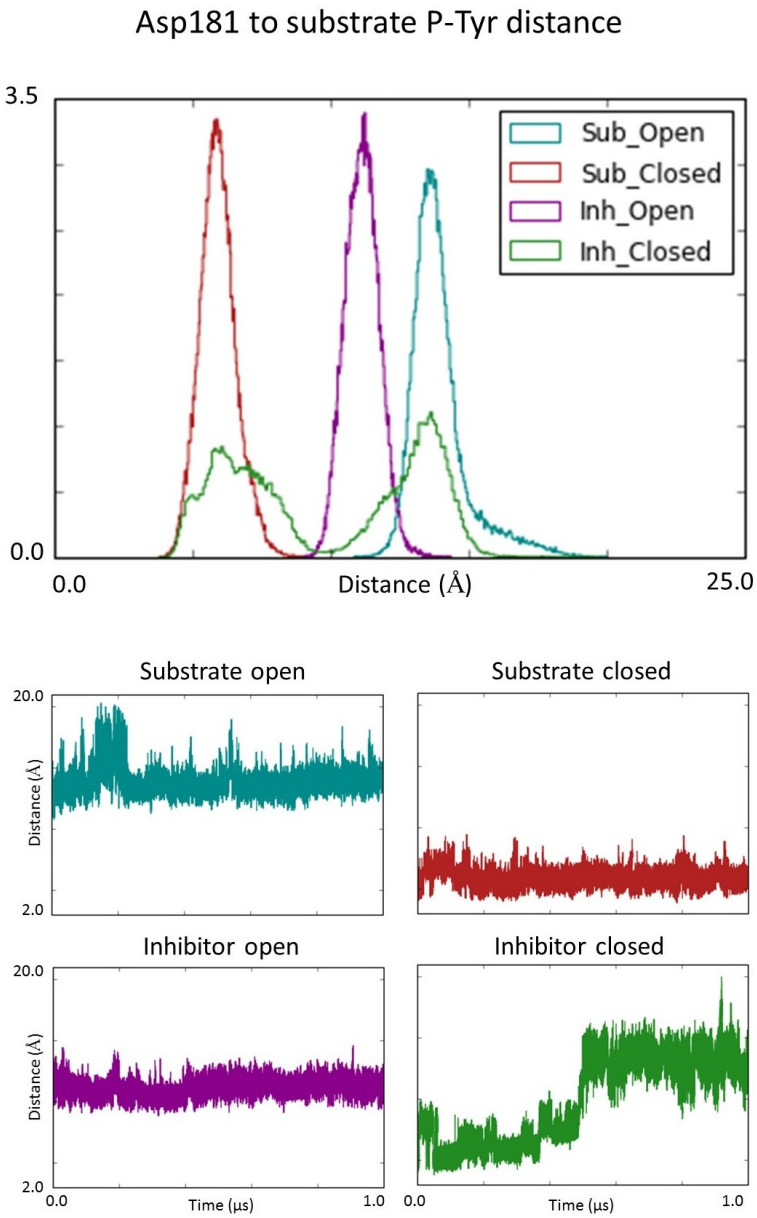


FIGURE 4.16: Distance from Asp181 to substrate P-Tyr. Teal: SO. Red: SC. Purple: IO. Green: IC.

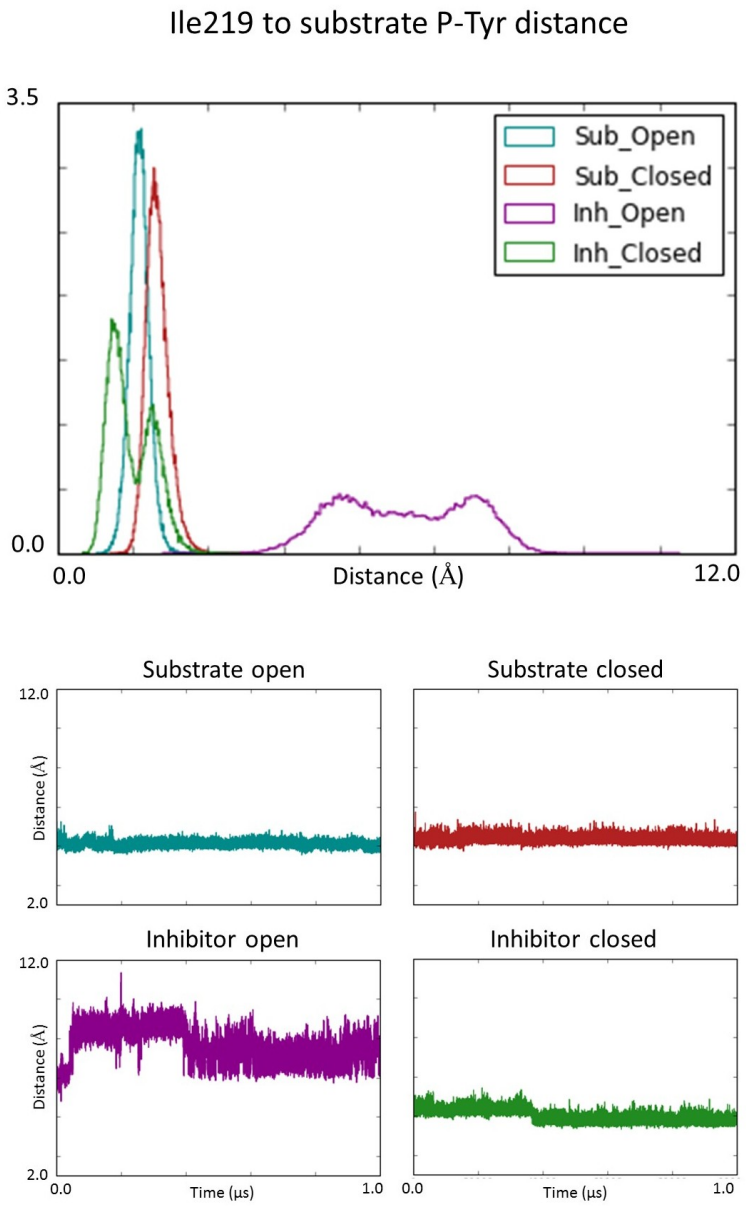


FIGURE 4.17: Distance from Ile219 to substrate P-Tyr. Teal: SO. Red: SC. Purple: IO. Green: IC.

4.3.6 PCA on Cα coordinates

The first and second eigenvectors correspond to 17.88% and 15.67% of the variance respectively. The sum of the first 10 eigenvectors accounts for 60.85% of the variance. The first principal component relates to both the WPD loop (residues Pro180-Val184) and the R-loop (residues Met114-Lys120) with a higher contribution to the variance from the R-loop. A per atom contribution to PC1 is shown in figure 4.18, along with distributions of PC1 for each system, and structures corresponding to the maximum and minimum values of PC1.

PC2 also relates to another motion of the WPD loop, and per atom contributions, distributions of PC2, and minimum and maximum structures are shown in figure 4.19. Plotting PC1 and PC2 as a 2D histogram, two states can be seen (figure 4.20(A)), one representing the open loop and the other the closed loop. Each of the four trajectory can then be plotted on this projection (figure 4.20(B)), which highlights that only the inhibitor bound simulation (green) which begins from the closed conformation, moves to the other, open state. This confirms the analysis discussed in 4.3.3.

For both PC1 and PC2, the JS divergence was computed in order to compare the distributions for each system. PC1 has higher JS divergence between the open and closed simulations, with maximum values of over 0.6. Substrate and inhibitor simulations starting from the closed conformation are similar, as are those starting from the open conformation, which can be seen with JS values of around 0.15 or less for both. PC2 results show the differences which were discussed previously. The substrate and inhibitor simulations beginning from the closed conformation show around the same JS divergence as the inhibitor closed to the substrate open. The largest JS divergence of PC2 is between the substrate open and substrate closed simulations, as there is only a very small overlap in these distributions.

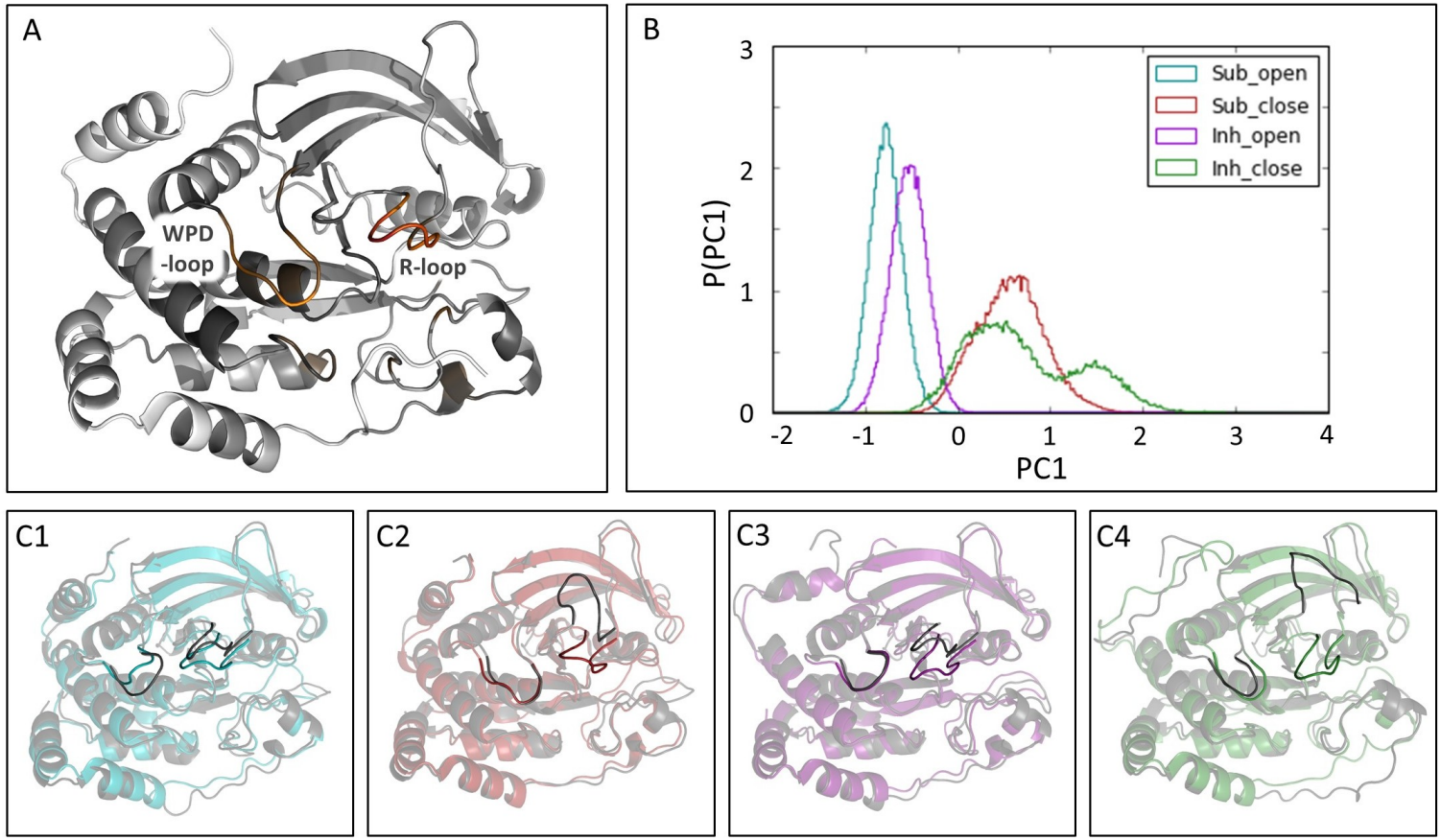


FIGURE 4.18: PC1 of four simulations: Teal: SO. Red: SC. Purple: IO. Green: IC. A: Per atom contribution to PC1. B: Distributions of PC1 for each system. C1-C4: Structures corresponding to minimum (grey) and maximum (colour) values of PC1.

FIGURE 4.19: PC2 of four simulations: Teal: SO. Red: SC. Purple: IO. Green: IC. A: Per atom contribution to PC2. B: Distributions of PC2 for each system. C1-C4: Structures corresponding to minimum (grey) and maximum (colour) values of PC2.

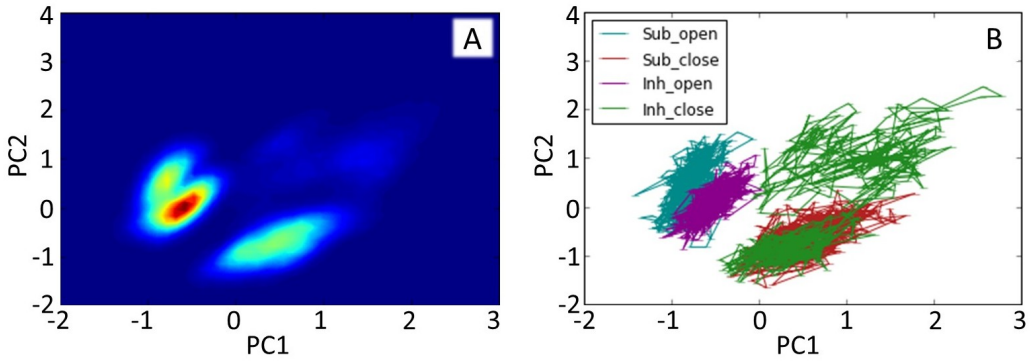


FIGURE 4.20: A: Distributions of PC1 plotted vs PC2, with increasing probability coloured blue-green-yellow-red. B: Same projection of PC1 vs PC2 with each trajectory superimposed. Teal: SO. Red: SC. Purple: IO. Green: IC.

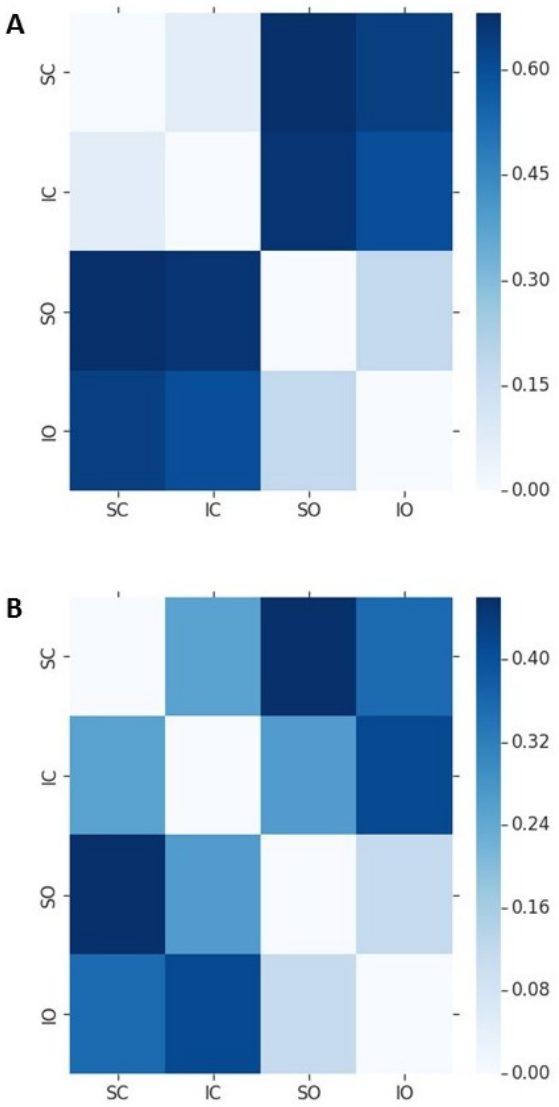


FIGURE 4.21: JS divergence for A: PC1, and B: PC2 for four systems. Clustering described in section 3.2.8, using ϵ=0.2 with 2 states.

4.3.7 MSM

As we have information for this system on both the active and inactive conformations, it would be useful to better understand the mechanism which causes the inhibitor to destabilise the closed conformation. The previous analysis gives some insights to the loop opening when inhibitor is bound, however we have not been able to sample the loop closing using equilibrium MD simulations. Furthermore, it is important to understand in more detail the mechanism which the inhibitor causes the loop to open and to confirm that this occurs in a statistically significant way.

To determine a possible pathway and intermediate states, an MSM was constructed, which uses four equilibrium MD simulations (substrate with open WPD loops; substrate with closed WPD loop; substrate and inhibitor with open WPD loop; substrate and inhibitor with closed WPD loop), and the equilibrium MD runs which were seeded from the steered MD for each system, as described in section 4.2.2.2.

The initial model constructed uses separate clustering for the substrate bound set, and the substrate with inhibitor bound set. This resulted in reasonably similar clustering however direct comparison is not possible. The second model allowed for clustering to be done on the entire dataset. In both cases, the "active" conformation has been defined as the closed WPD loop (using the RMSD of both backbone and sidechain atoms) with the substrate at shorter distances to the active site cysteine residue. The "inactive" conformation is defined as the open WPD loop with the substrate at longer distances.

4.3.7.1 Initial MSM model

A range of implied timescales were computed and the results can be seen in figure 4.22 for each system. From the implied timescales, a lagtime of 3000 steps (30 ns) was selected to construct the MSM.

Plots highlighting the structure of a three macrostate model are shown in figure 4.23 for both the substrate only, and the substrate with inhibitor (FRJ) simulations. Clustering in this case is done separately for each set of

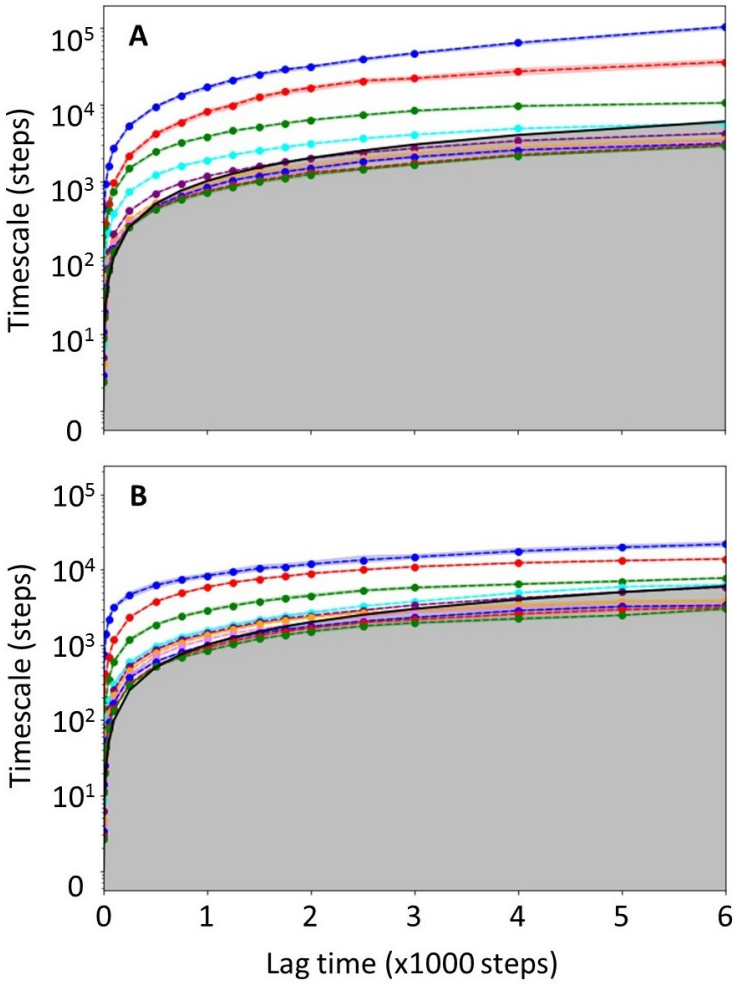


FIGURE 4.22: Implied timescale for substrate and inhibitor simulations based on each set clustered separately, with 100 clusters.

trajectories. As a result, timescales and populations of each state cannot be directly compared, however as the input dimensions are easy to interpret, comparison of the relative populations and relative timescales are still valid. This shows that the most active state (orange) is much more likely with no inhibitor bound (9.83 %), than with inhibitor bound (0.63 %). Also when inhibitor is bound almost 90 % is assigned to the most inactive state (magenta). This can be explained by considering the rates of transition between each state. These are computed as the mean first passage time (MFPTs) between macrostates which are calculated as a weighted average of MFPTs between each pair of microstates. Uncertainties are estimated as standard deviations of the mean. This shows that both the magenta and teal states have longer relative timescales (i.e. the ratio of the transition timescales in and out of a state) to transition to the active state (orange) when inhibitor is bound.

In both cases, three states are defined by the different coloured cluster centres: magenta (open WPD loop with substrate at longer distances to active site); teal (open WPD loop with substrate at shorter distances to active site); and orange (closed WPD loop with substrate at shorter distances to active site). With no inhibitor bound, the clusters for the magenta state reach a maximum substrate distance of around 8 Å. However in the inhibitor bound simulations, the maximum distance is much larger. This can be seen in more detail in figure 4.24, where structures for each state shown in colour represent the average structure for each macrostate, based on selection of 1000 structures from each state. With only substrate bound, the average structure in the magenta state has a substrate to active site distance of 4.1 Å however when inhibitor is also bound the average structure has an average substrate distance of 6.7 Å.

The selection of three macrostates to construct the MSM was validated by carrying out Chapman-Kolmogorow (CK) test, to confirm that the the Markov property holds. The results of this test are seen in figures 4.25 for the substrate only set, and 4.26 for the inhibitor bound set.

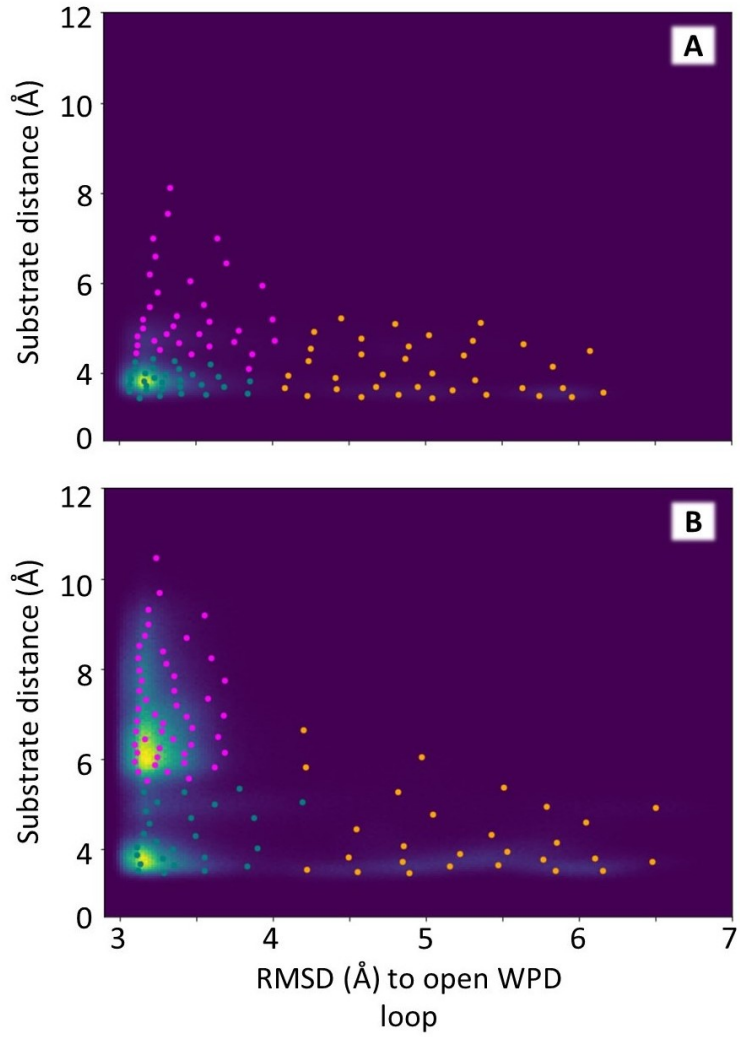


FIGURE 4.23: Clustering using 100 k-means clusters for each set done separately, with A: Substrate only set and B: Substrate with inhibitor FRJ set. Colours of clustercenters correspond to macrostate assigned to.

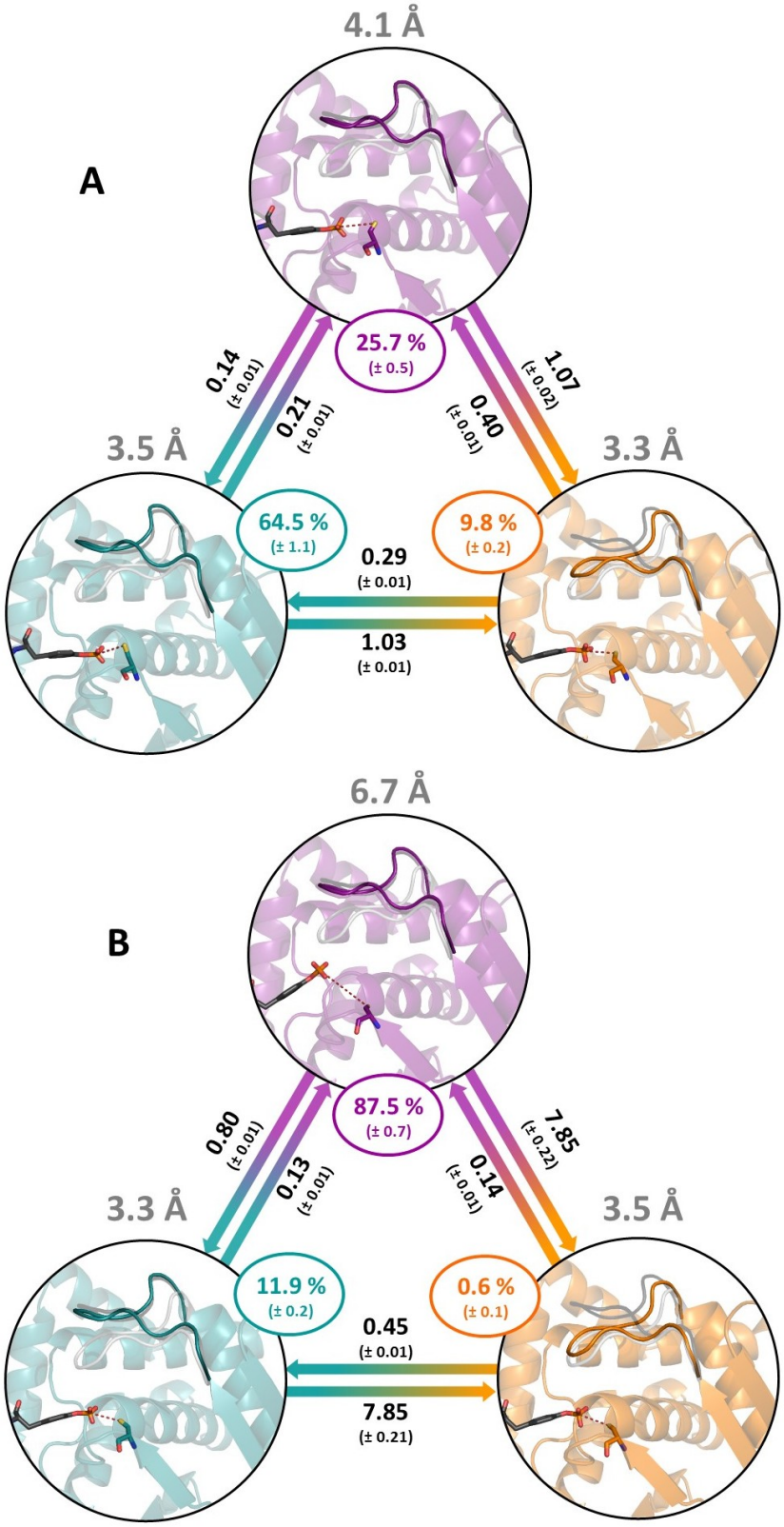


FIGURE 4.24: Three macrostates defined based on colouring from figure 4.23. A: Substrate bound simulations. B: inhibitor bound simulations. Transition timescales are in units of μs. Distances noted are the average distance of active site Cys(S) to substrate P-Tyr(P) for each state.

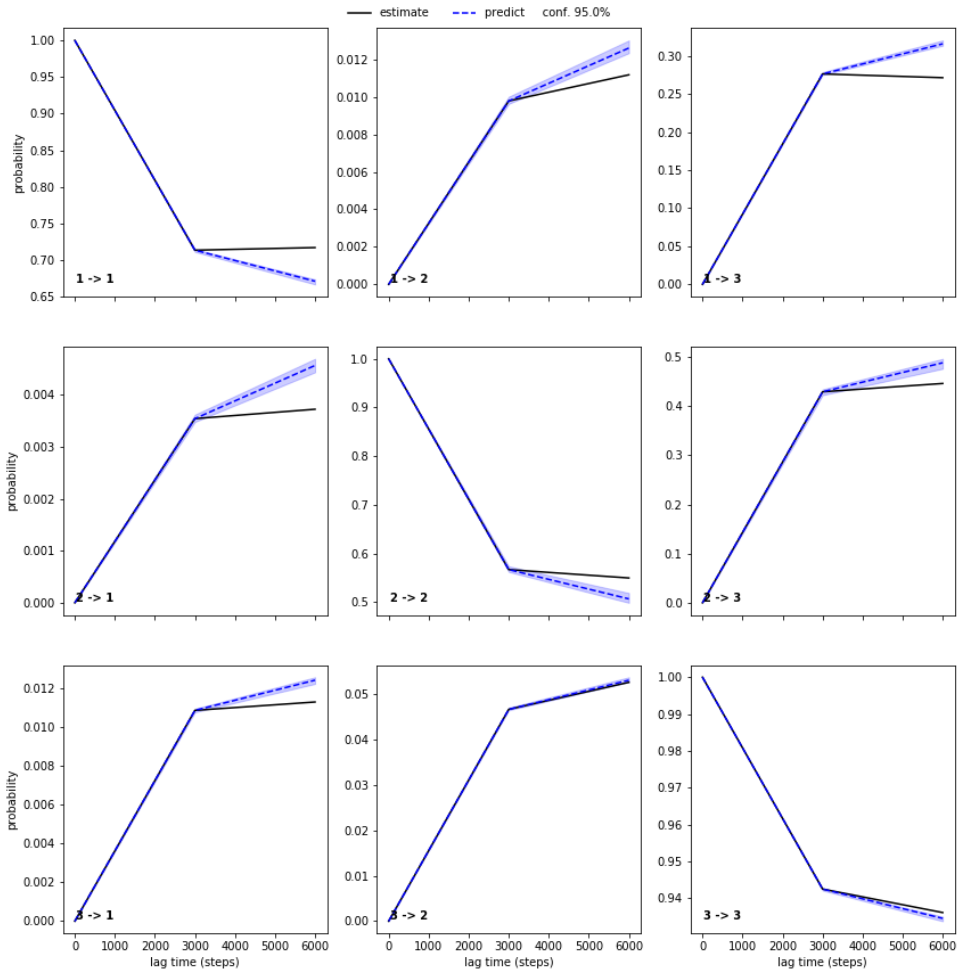


FIGURE 4.25: Chapman-Kolmogorow (CK) test for substrate model with separate clustering.

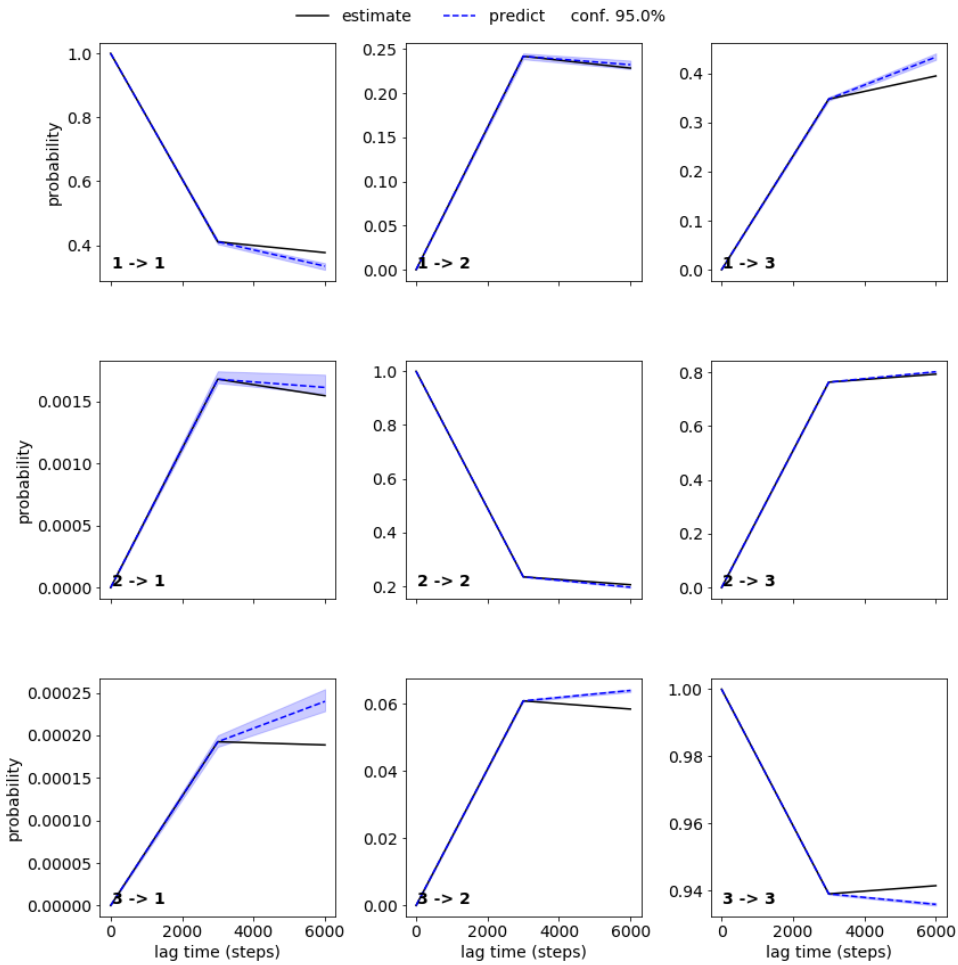


FIGURE 4.26: Chapman-Kolmogorow (CK) test for inhibitor model with separate clustering.

4.3.7.2 Improved MSM model

A second MSM was generated in order to attempt to cluster the entire dataset and project both the substrate, and the inhibitor bound simulation sets onto the same cluster centres. In figure 4.27, the implied timescales for a range of lag time τ are shown. Considering the result for both the substrate and inhibitor plots, a lag time of 2000 steps (20 ns) was selected to build the MSM. In the implied timescale plot for the substrate set, it is possible to resolve three separate processes, whereas in the inhibitor set this is likely only two. Timescales are similar but slightly slower in the substrate data set (converging between 10^3 to 10^4 steps) compared to the inhibitor set (converging around 10^3).

In this instance, instead of using the hidden markov model to obtain macrostates, PCCA was used. This allowed for three macrostates to be defined for each system, which correspond to the most active state (orange: WPD loop closed, substrate at short distances), an intermediate state (teal: WPD loop open or closed, substrate at short-medium distances), and the inactive conformation (magenta: WPD loop open, substrate at longer distances). This now allows direct comparison of the substrate, and substrate with inhibitor simulations.

A coarse grain description of the microstates shows variation in the intermediate (teal) state when comparing both the substrate and inhibitor sets (figure 4.28). For the substrate bound simulations only 18 of the 100 clusters are assigned to the intermediate (teal) state. In contrast, the inhibitor bound simulation set assigns far more clusters to this state.

In figure 4.29 the conformations for each state are highlighted. Structures for each state represent the average structure for each macrostate, based on selection of 100k structures from each state. For the inhibitor bound simulations, over 95 % of structures as assigned to the inactive (magenta) conformation. While for the substrate set this is only around 5 %. The conformation of the intermediate (teal) conformation varies between each system. As the structures used to illustrate this model are averages of structures assigned to each state, these vary due to the assignment of clusters shown in figure 4.28. The substrate set has far more clusters assigned to the teal

macrostate, and as such the average structures are obtained from a different number of cluster centers.

It can be seen from the transition rates between states shown in figure 4.29, that the transitions into the active (orange) state from both the inactive (magenta) and intermediate (teal) states are slower when in inhibitor is bound, by around a factor of 8. In addition, the timescale to move from the active to inactive states directly is faster when inhibitor is bound. As was discussed in section 4.3.5 and in figure 4.15, the substrate remains at reasonably close distances to the active site even with the open WPD loop conformation, only when no inhibitor is bound. This then allows for the higher population of the intermediate (teal) state for the substrate simulation set.

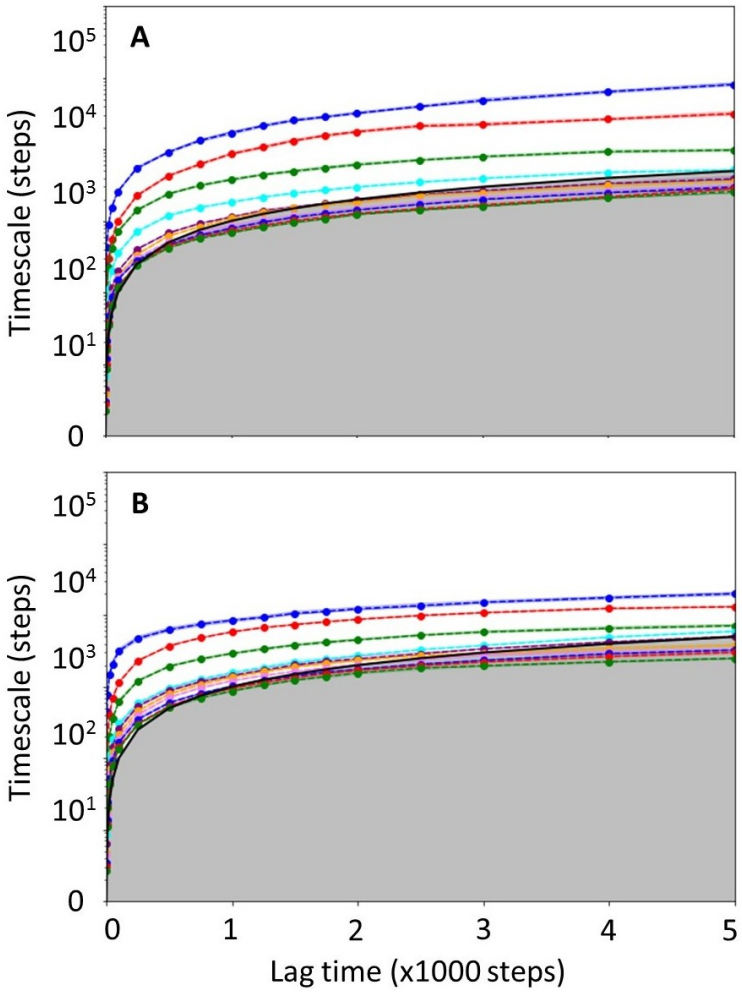


FIGURE 4.27: Implied timescale for A: substrate and B: inhibitor simulations based on combined clustering, with 100 clusters.

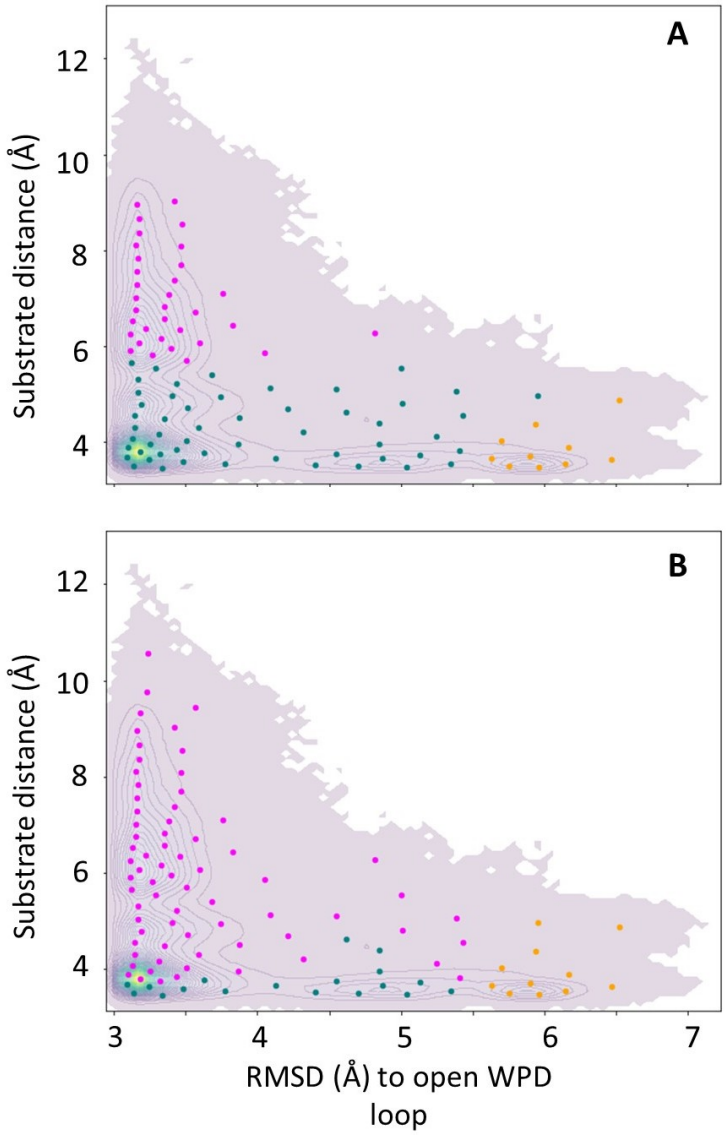


FIGURE 4.28: Clustering using 100 k-means clusters, with A: Substrate only set and B: Substrate with inhibitor FRJ set. Colours of clustercenters correspond to macrostate assigned to.

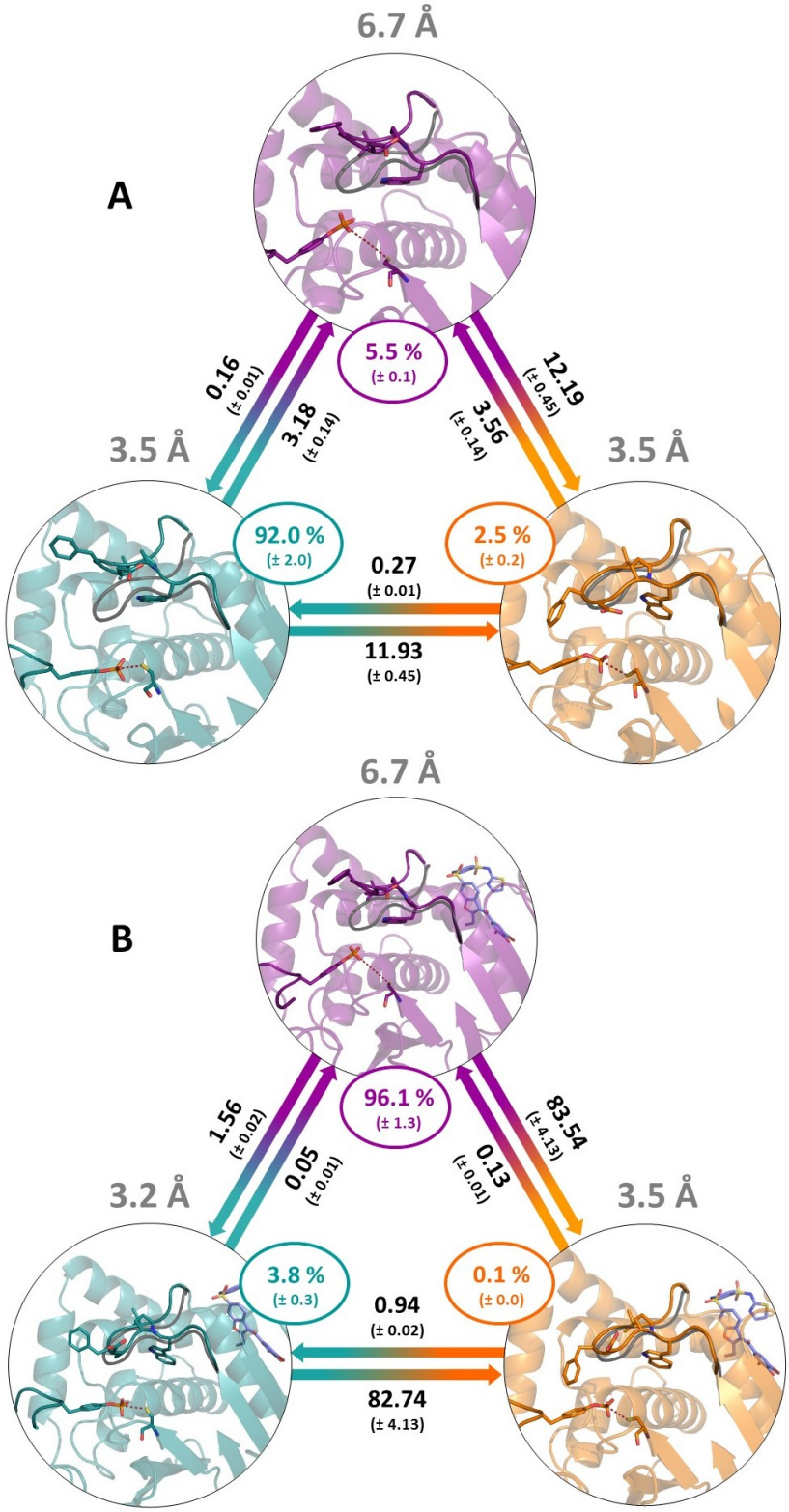


FIGURE 4.29: Three macrostates defined based on colouring from figure 4.28. A: Substrate bound simulations. B: inhibitor bound simulations. Transition timescales are in units of μs. Distances noted are the average distance of active site Cys(S) to substrate P-Tyr(P) for each state.

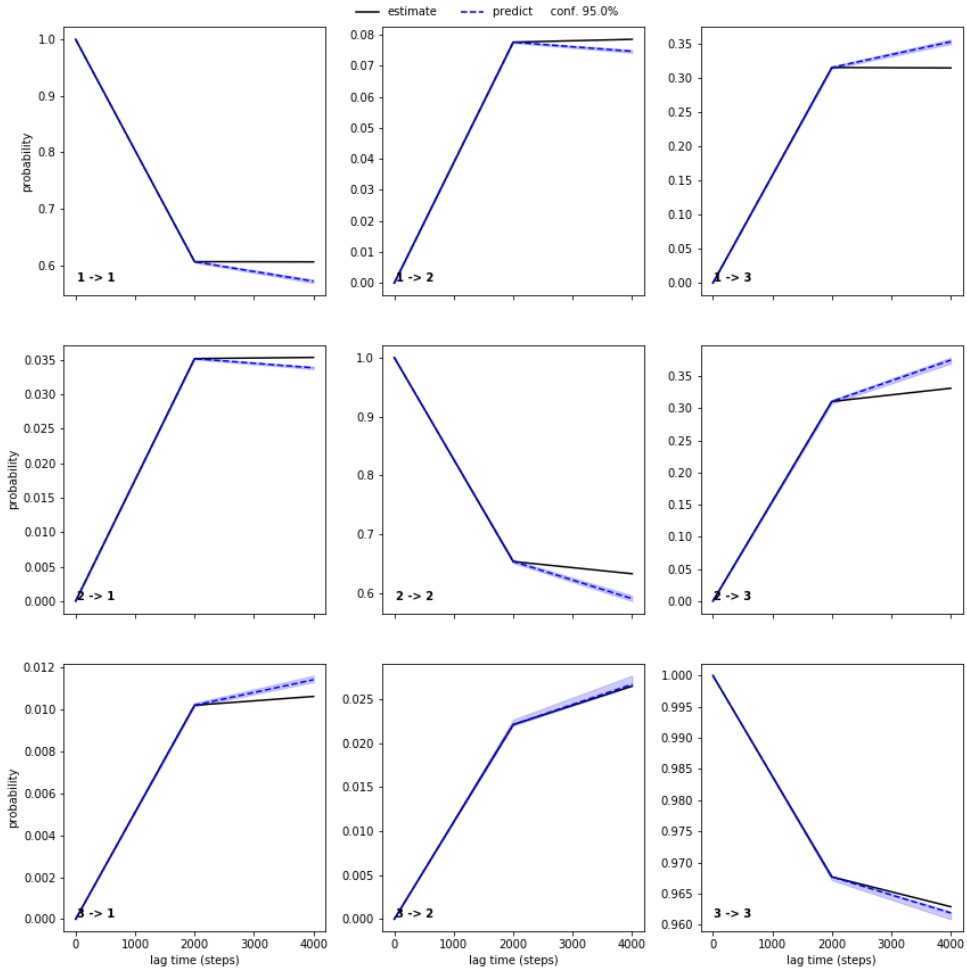


FIGURE 4.30: Chapman-Kolmogorow (CK) test for substrate model with combined clustering.

Again the Chapman-Kolmogorow (CK) test was used to determine if the three state model was suitable and results are shown in figures 4.30 and 4.31. In both this model and the previous model, the CK test confirms that the processes obey Markovianity.

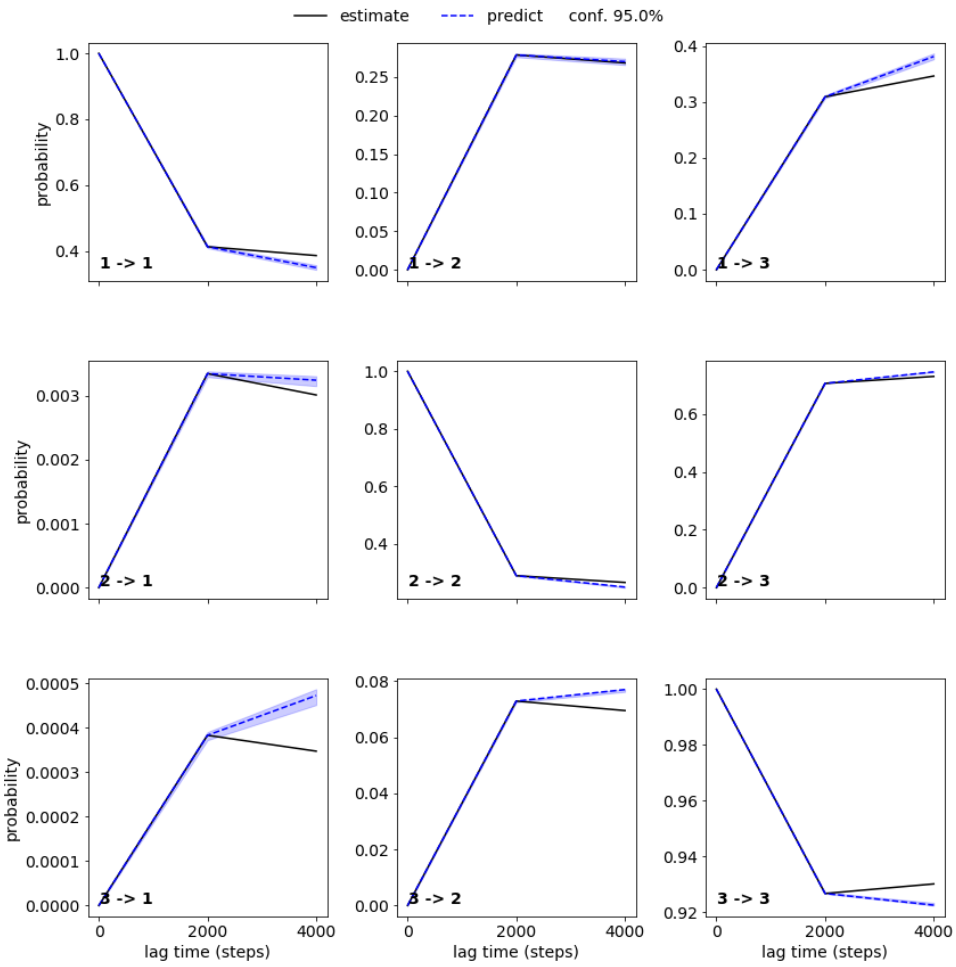


FIGURE 4.31: Chapman-Kolmogorow (CK) test for inhibitor model with combined clustering.

4.4 Discussion

Equilibrium MD simulations of PTP1B starting from both the open and closed WPD loop conformations, with and without allosteric inhibitor, have given some insights in the motions of the WPD loop. In each of the simulations with two different allosteric inhibitors (FRJ and D0P), the closed conformation was not stable, and the loop opens on a timescale of a few hundred nanoseconds. Starting from the open conformation, loop closing was not observed in any of the simulations, which was expected as it is believed that this transition occurs on the order of μs to ms. Analysis of RMSD values higlighted that the open conformation does not seem to vary between the inhibitor bound and inhibitor free simulations.

The two MSM models constructed both qualitatively show the same result: that the inhibitor bound set has the highest probability of occupying the inactive (open WPD loop; long substrate distance) state. Comparison of both models by comparing the ITs and the CK test plots does not suggest that one model is better than the other, however as the clustering was done for the second model using the entire data set, then the resulting populatoins of states, and transition timescales for both the substrate and inhibitor systems can be compared directly. In both models, transitions into the most active state (closed WPD loop; short substrate distance) when inhibitor is bound are much slower than when no inhibitor is bound, and populations of the most active state are higher in the substrate only set compared to the inhibitor set.

Chapter 5

Conclusions

In this thesis, molecular dynamics simulations combined with an information theory based analysis have provided key insights into the mechanism of allosteric regulation of PDK1 and PTP1B. The results highlighted the need for tailored analysis methods, on which different metrics are required to capture the differences between activated and inhibited conformations.

In the case of PDK1, movements of the activation loop were highlighted using both the KL divergence of dihedral angles, and with Cα coordinate PCA. The KL analysis showed largest variations were in residues at the hinge region of the activation loop, while PCA confirmed this motion to show the largest variance in atomic positions. Furthermore, the PCA analysis allowed to identify distinct conformations of the activation loop between the complexes of activator and inhibitor molecules. Specifically, in the complex with inhibitor bound, it was found that the loop cannot adopt a conformation that allows the substrate to be positioned within a reactive distance to ATP. Calculating the JS divergence with application of spectral clustering allowed for large sets of compounds to be compared easily, using different metrics such as PCA, or distances. In addition, MI allowed to correlate the activation loop motion with the interaction energy of ATP. The MI values obtained are relatively small, however allostery may only be a result of very subtle changes between activated and inhibited conformations. As the results are adjusted for noise, these values are still sufficient to confirm correlation between the ATP interactions and the loop motion. Hence the computational framework developed allowed to provide previously unknown structural rational for the allosteric effects triggered by small molecules in

PDK1. Attempts to rank compounds according to level of activity proved to be more difficult, as while differences were still seen for most descriptors when comparing activations to the inhibitor bound simulation, the degree of activation did not seem to show any trend in the descriptors caculated.

For PTP1B, based on the available literature on the regulatory role of the WPD loop, enhanced sampling methods were used to capture the loop closing process. These simulations shed light into the molecular determinants underpinning the regulation process, challenging the previous hypothesis of side chains causing mechanical hindrance to the movement of the loop. In contrast, results provide a thermodynamical picture that suggest that allosteric inhibitors destabilise the catalytically active comformation of the WPD loop with respect to inhibitor free complexes, but closing of the loop is still possible in inhibitor bound complexes. This can be seen from the MSM results, as there is still some population in the active (closed WPD loop) conformation, however the population of this state is low for inhibitor bound PTP1B when compared to the population when only substrate is bound.

Equilibrium MD analysis for PTP1B found that the two different allosteric inhibitors caused the loop to open within a few hundred nanoseconds when simulations were started from the closed WPD loop conformation. Simulations which were started from the open WPD loop conformation did not show significant differences in the RMSD of the WPD loop between inhibitor bound, and inhibitor free simulations. Calculation of distances showed that only the inhibitor bound, the open WPD loop simulation allowed the substrate P-Tyr to move further away from the active site. For the substrate bound, open loop simulation, the substrate P-Tyr still remains reasonably close to the active site. This suggests that perhaps the difference in stabilisation of the substrate with or without inhibitor is a factor, before the loop can close. PCA of set of four simulations with inhibitor FRJ and without inhibitor highlight the differences in conformation of the WPD loop, and also the R-loop. In the closed WPD loop conformation, the variance in this loop motion is much larger, than when the WPD loop is open.

Future work on this project should aim to extend the MI analysis with energy decomposition. This will lead to a better understanding of specific

interactions that allosteric ligands established and that that lead to structural differences in distant sites to the ligand binding pocket. Prediction of these patterns will aid on the design design of new allosteric drug candidates. Furthermore, combining interaction energy MI with the MSM analysis will unravel the molecular causes of the stabilisation of individual states.

Appendix A

Phosphoinositide-dependent kinase-1: PDK1

A.1 Specific distance figures

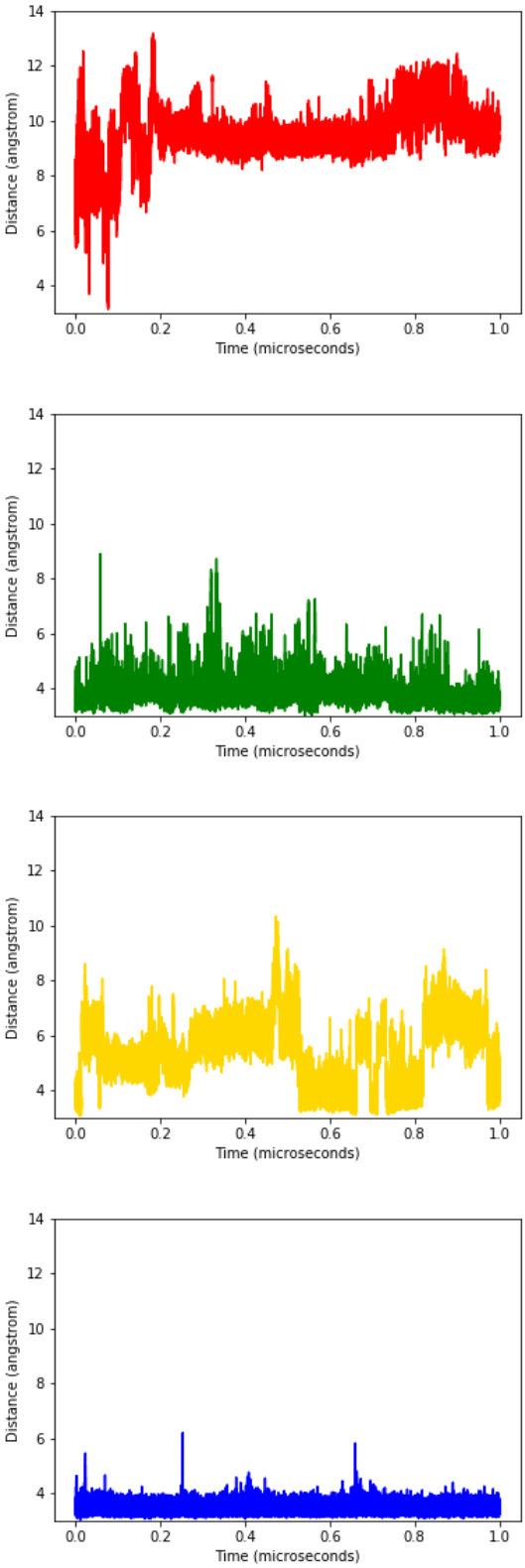


FIGURE A.1: Distance per snapshot of substrate peptide Thr to γ-phosphate of ATP distance for the four original simulations completed.

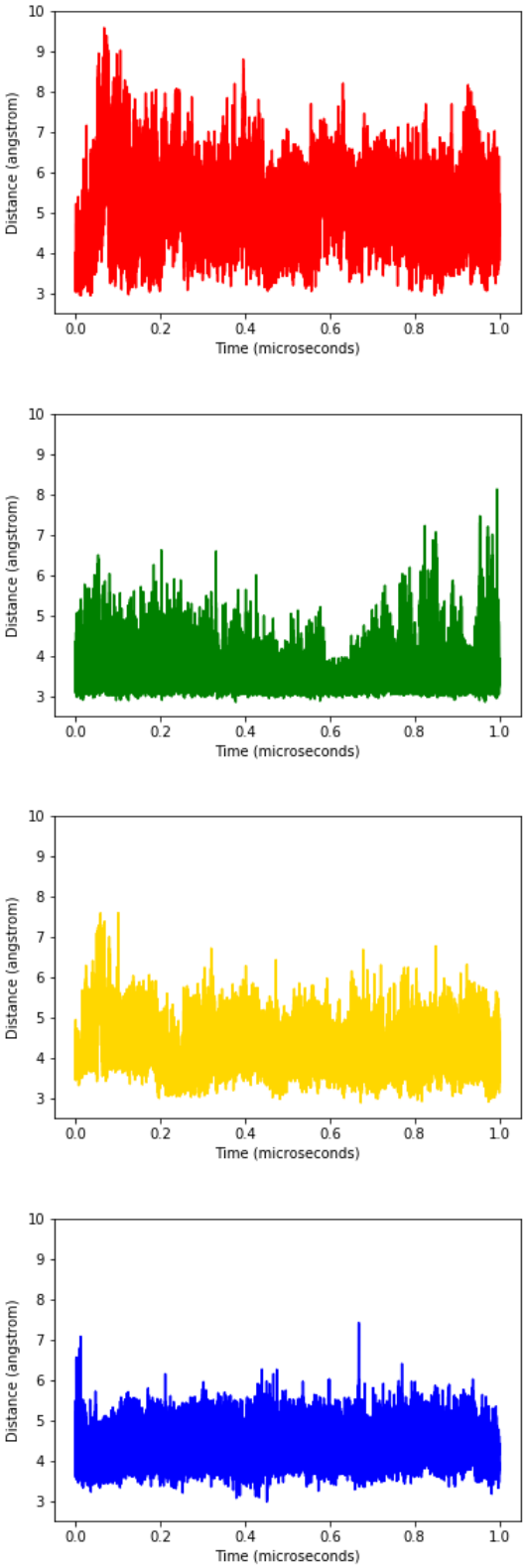


FIGURE A.2: Distance per snapshot of Lys39 to Glu58 distance for the four original simulations completed.

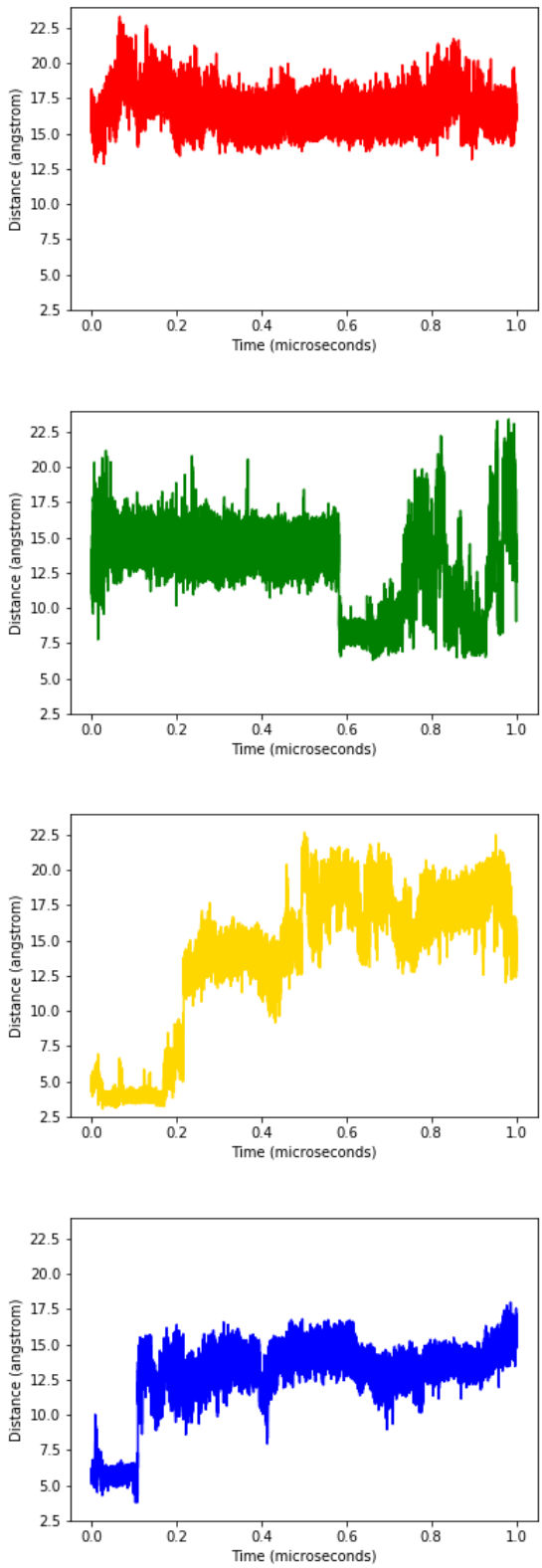


FIGURE A.3: Distance per snapshot of Tyr54 to Glu58 distance to γ-phosphate of ATP for the four original simulations completed.

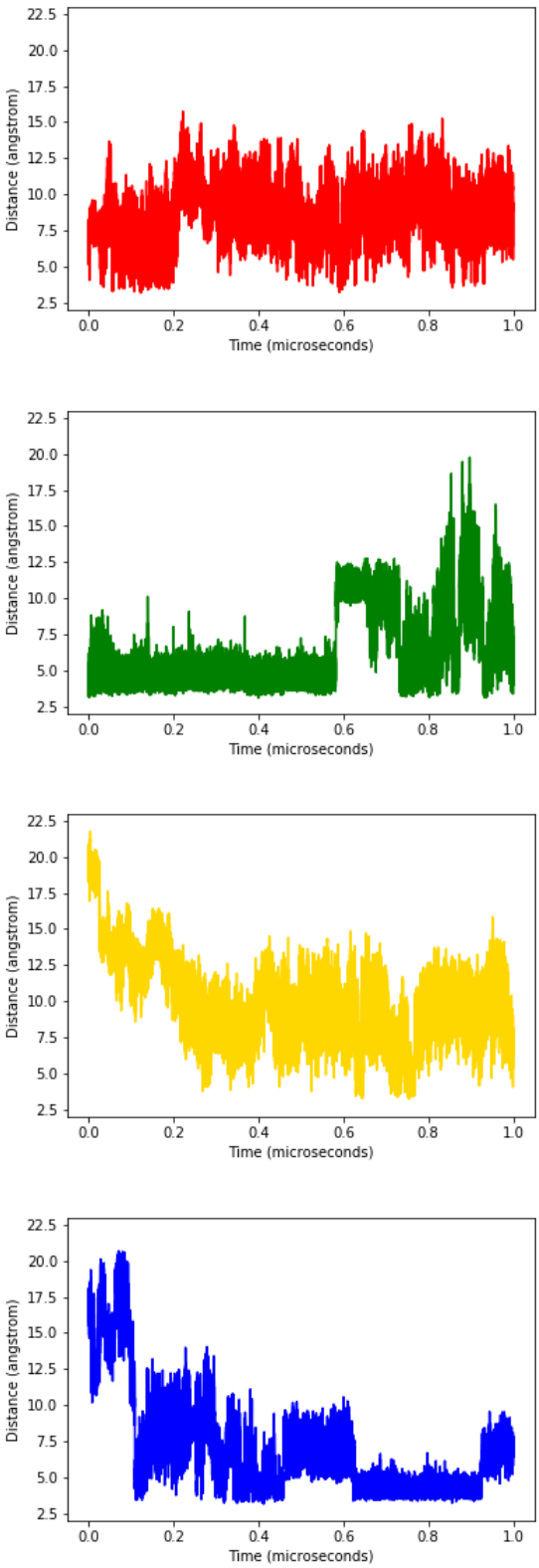


FIGURE A.4: Distance per snapshot of Tyr54 to phosphoserine distance to γ-phosphate of ATP for the four original simulations completed.

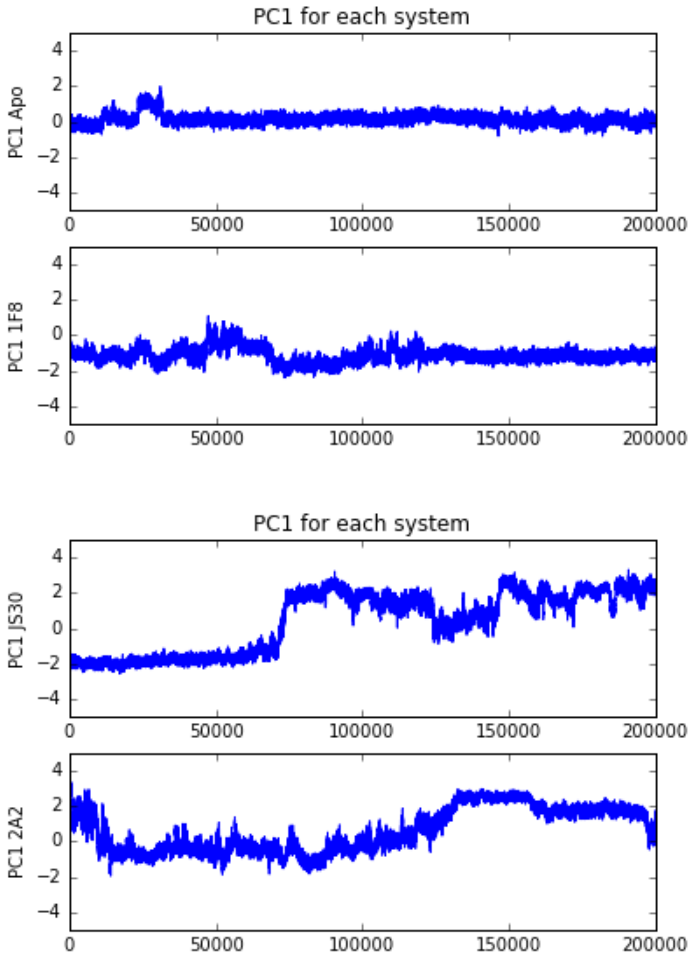


FIGURE A.5: PC1 value per snapshot for the four original simulations completed.

A.2 PCA figures

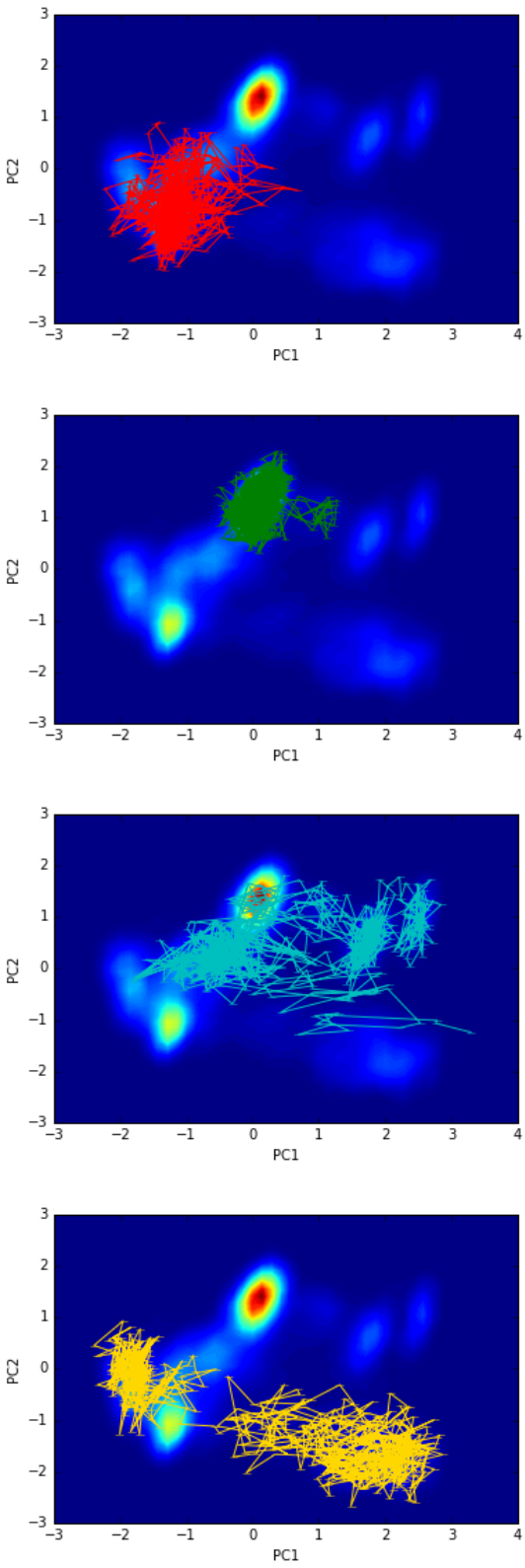


FIGURE A.6: PC1 vs PC2 2D distribution, with individual trajectories superimposed showing 1 every 300 snapshots.

A.3 MI testing

APO	**MI**	**MI(rand)**	**MI(corr)**
AB	0.985	0.112	0.872
AC	0.134	0.114	0.020
BC	0.142	0.114	0.028
1F8	**MI**	**MI(rand)**	**MI(corr)**
AB	0.966	0.134	0.833
AC	0.139	0.120	0.019
BC	0.150	0.120	0.031
2A2	**MI**	**MI(rand)**	**MI(corr)**
AB	1.372	0.187	1.184
AC	0.138	0.131	0.007
BC	0.145	0.133	0.012
JS30	**MI**	**MI(rand)**	**MI(corr)**
AB	1.243	0.175	1.069
AC	0.228	0.153	0.075
BC	0.233	0.146	0.087

TABLE A.1: Full testing values for MI computed between distances A, B and C using 200,000 snapshots and 300 bins.

APO	**MI**	**MI(rand)**	**MI(corr)**
A	0.033	0.030	0.003
B	0.031	0.029	0.002
C	0.031	0.030	0.001
ORX	**MI**	**MI(rand)**	**MI(corr)**
A	0.207	0.035	0.172
B	0.129	0.036	0.094
C	0.070	0.035	0.035
ORZ	**MI**	**MI(rand)**	**MI(corr)**
A	0.498	0.035	0.463
B	0.423	0.036	0.387
C	0.047	0.036	0.011
OTU	**MI**	**MI(rand)**	**MI(corr)**
A	0.161	0.036	-0.016
B	0.136	0.037	-0.009
C	0.073	0.035	-0.035

TABLE A.2: Full testing values for MI computed between ATP interaction energy, and distances A, B or C using 40,000 snapshots and 60 bins.

APO	**MI**	**MI(rand)**	**MI(corr)**
A	0.340	0.031	0.310
B	0.449	0.030	0.419
C	0.059	0.029	0.030
ORX	**MI**	**MI(rand)**	**MI(corr)**
A	0.488	0.033	0.455
B	0.493	0.032	0.461
C	0.050	0.032	0.018
ORZ	**MI**	**MI(rand)**	**MI(corr)**
A	1.067	0.040	1.027
B	1.204	0.041	1.164
C	0.042	0.033	0.008
OTU	**MI**	**MI(rand)**	**MI(corr)**
A	1.290	0.037	1.253
B	1.123	0.040	1.083
C	0.155	0.036	0.119

TABLE A.3: Full testing values for MI computed between PC1, and distances A, B or C using 40,000 snapshots and 60 bins.

Appendix B

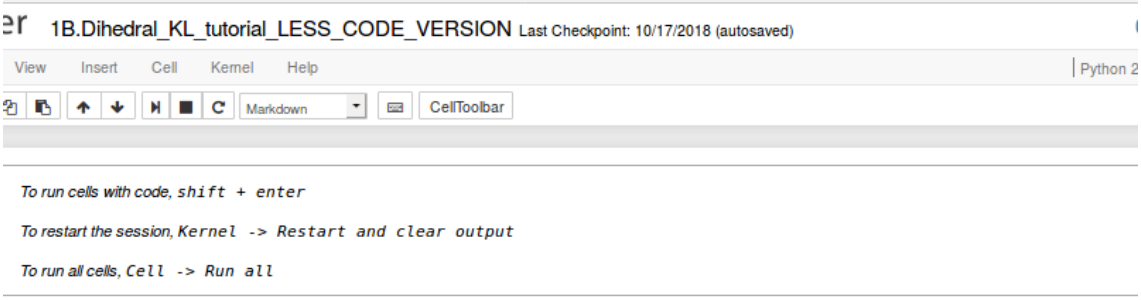
PDK1 analysis scripts

A brief summary of the scripts available on GitHub to carry out the PCA, KL and MI analysis is provided below. A Jupyter notebook tutorial can be found in the GitHub repository [121].

The tutorials include an overview of the theory, method, and explain how to format input trajectories and directory format, prior to running the scripts. Short trajectories of PDK1 are provided in order to run the notebooks directly. Further scripts to run the analysis from the command line are also provided.

The first notebook shown in figure B.1 details the KL divergence of torsional angles. This includes details on how KL is calculated, and explains the overall workflow as shown in figure 3.13. It is then possible to launch a Pymol session with a script provided, which will assign KL values to the B-factor column of the PDB file, and then visualise these as a colour scale on the structure in Pymol. The Pymol session will include six structures. 'KL_backbone': summed ψ and ϕ; 'KL_sidechain': summed $\chi 1$ and $\chi 2$; and then each ψ, ϕ, $\chi 1$ and $\chi 2$ shown separately.

The notebook to run the PCA analysis shown in B.2 again details some background on PCA, and information on the output of the analysis. The PCA analysis is run, and the resulting output can be loaded into pymol to visualise both the per atom contribution to the first and second principal component, and the structures representing the minimum and maximum values of PC1 and PC2 for the input trajectories.



Dihedral Kullback-Leibler (KL) divergence

Introduction

Some details on trajectory/topology input

Simulation input

To run these scripts, you should have two (or more) different trajectories for the same protein, with different effectors bound. In these examples, the simulations have different allosteric ligands.

Topology

It is useful that the data is first processed to ensure the same residues are present. This script will calculate psi/phi/chi in the order which they are in the topology, and assign using the index of the Cα atom for that torsion. Therefore having topology with the same atom indices ensures that for each system the output is the same.

Trajectory

All of these scripts will take input trajectories which are accepted by mdtraj.

There are some basic scripts in the folder **`Scripts/Traj_processing_scripts`** for processing trajectory and topology using parmed or cpptraj. Remove any protein residues not common to both trajectory and also you can remove water as it will not be needed.

As long as you have the same sequence of residues for the protein in each case, there is no need to remove any substrate, ligand or ions, but it is useful to do also this now if you are also going to carry out a PCA.

Kullback-Leibler divergence (KL)

Usually for each KL calculation, two different simulations are compared. These could be a simulation with no ligand bound and one with a ligand. Or two simulations with different ligands (i.e. activating or inhibiting ligands).

Also in the scripts folder is a script to create two smaller trajectory from one larger one. It is useful to calculate the KL divergence of one system relative to itself, using two different (or one separated) trajectories, in order to take account of noise.

Kullback-Leibler (KL) Divergence

KL divergence is defined as:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

In this case, distribution P will be the data obtained from one simulation *(i.e. protein with activator bound)* and distribution Q will be that obtained from another simulation *(i.e. protein with inhibitor bound)*. This is computed over a range of *i* bins.

FIGURE B.1: Section of the tutorial to run KL divergence of torsional angles available on GitHub [121].

2B.PCA_tutorial_CA_coordinates_LESS_CODE_VERSION (unsaved changes)

View Insert Cell Kernel Help Python

Markdown CellToolbar

Some background on PCA

From a set of data X_i which we obtain from the trajectory, we compute the covariance matrix:

$$C = (X - \mu)^T (X - \mu)$$

and solve the eigenvalue problem:

$$Cr_i = \sigma_i r_i$$

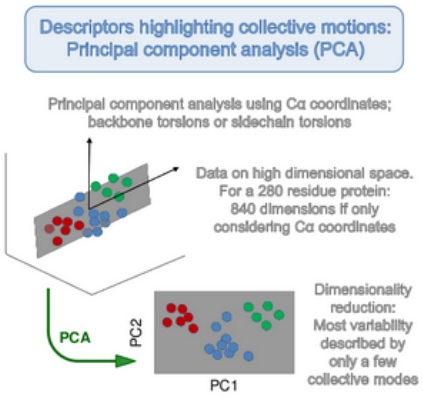
r_i are the principal components and σ_i are their respective variances.

The input data X_i can be something like Cα coordinates, or backbone or sidechain torsions.

In order to compare different simuluations, we do the dimensionality reduction on all available trajectories of the system. In this case, we have 4 different simulations, and will input all into the PCA together. We can then compare the highest variance motions between the different systems.

Reduction of dimensionality

This therefore allows a large dimensional dataset (i.e. all Cα coordinates in x, y and z directions) to be reduced into a smaller number of dimensions, where the new set of dimensions should still account for a large amount of the variance.



The output of this is a set of principal components, with Principal Component 1 (PC1) having the highest variance, and subsequent PC's having decreasing variance.

As a guideline, we usually calculate the first 10 principal components, and we can check how much of the variance these first 10 PCs account for.

The following script selects a subset of residues (it excludes the terminal regions of the model protein) and carries out a Cα coordinate PCA.

Output from the PCA

The idea is to also use the output from the PCA in calculations of MI (Mutual Infomrmation) or KL (Kullback-Leibler) Divergence, therefore we need to output the values of PC per snapshot for each system, and distributions of these values.

The script will output several things:

- Frames corresponding to the minimum and maximum values of PC1 and PC2 for each system.
- Per atom contribution to PC1 and PC2.
- Per snapshot value of PC1 and PC2 for each system.
- Distribution of values of PC1 and PC2 for each system.

FIGURE B.2: Section of the tutorial to run PCA analysis available on GitHub [121].

Appendix C

Protein tyrosine phosphatase 1B: PTP1B

C.1 Loop RMSD figures using larger number of residues.

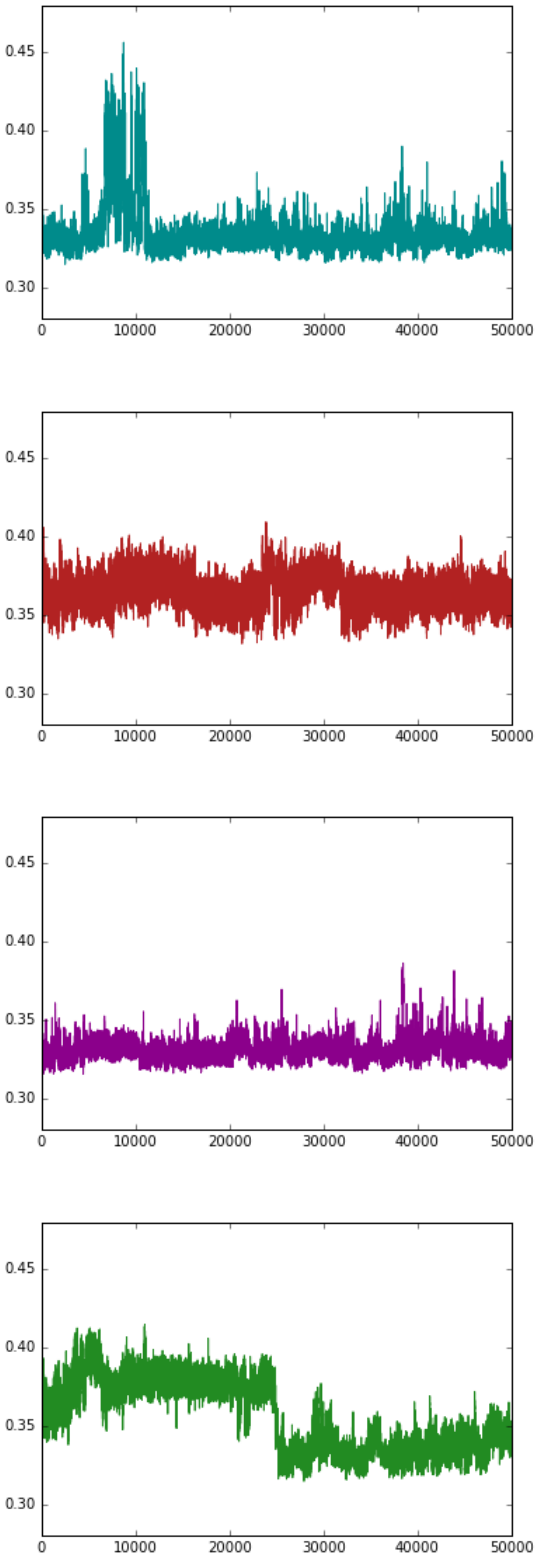


FIGURE C.1: RMSD for residues Thr177-Glu186 relative to open loop. Teal: substrate open. Red: substrate closed. Purple: inhibitor open. Green: inhibitor closed.

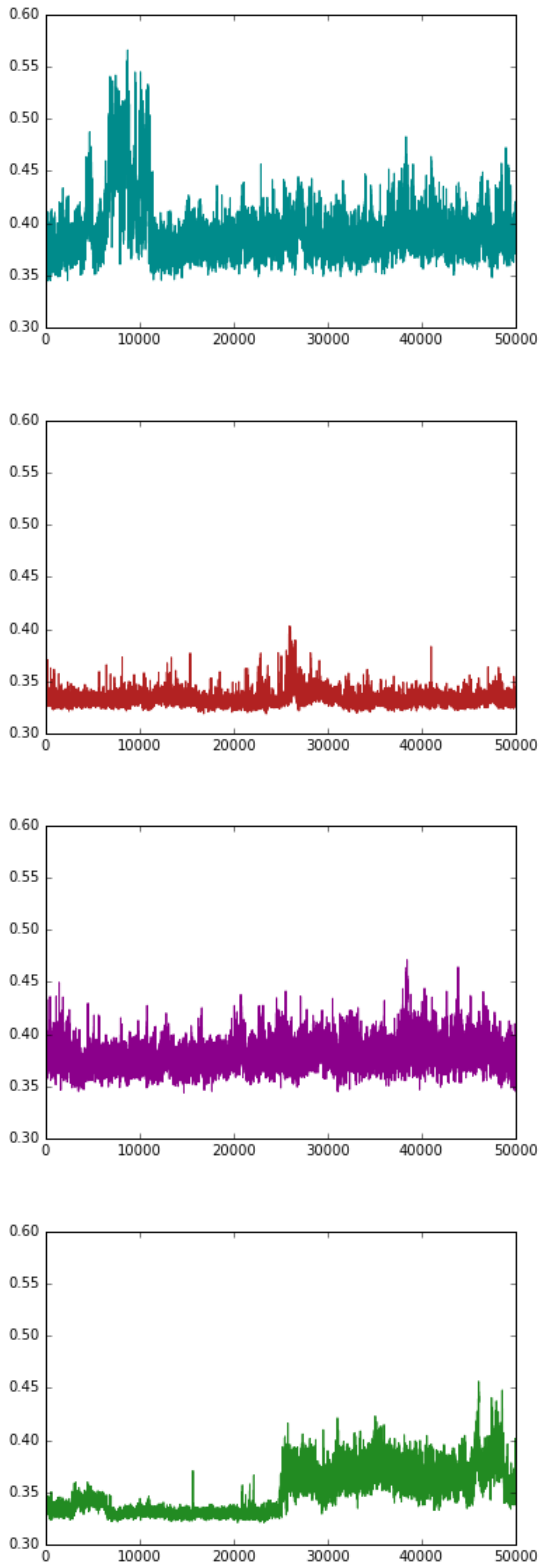


FIGURE C.2: RMSD for residues Thr177-Glu186 relative to closed loop. Teal: substrate open. Red: substrate closed. Purple: inhibitor open. Green: inhibitor closed.

C.2 PCA

Plots showing value of PC1 and PC2 per snapshot for four simulations.

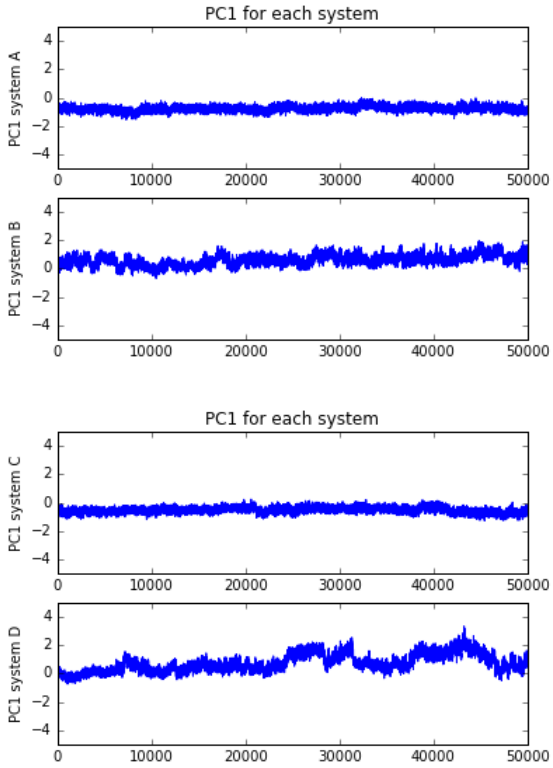


FIGURE C.3: PC1 value per snapshot for A: substrate open; B: substrate closed; C: inhibitor open; and D: inhibitor closed.

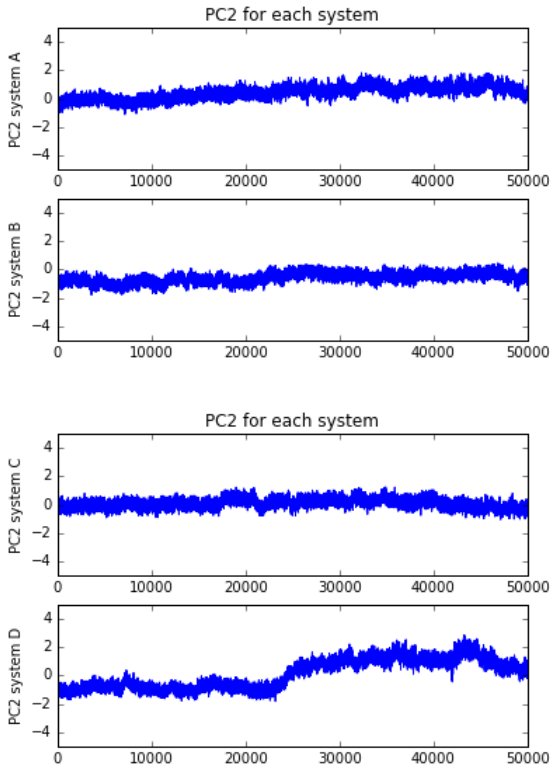


FIGURE C.4: PC2 value per snapshot for A: substrate open; B: substrate closed; C: inhibitor open; and D: inhibitor closed.

Appendix D

Presentations and posters

D.1 Oral and Poster presentations

D.1.1 Oral presentations

- CCP5 Summer School (Lancaster University, July 2016)
- Joseph Black conference (The University of Edinburgh, 2018)
- Astbury Conversation (Flash presentation: University of Leeds, 2018)
- MGMS Young Modellers' Forum (University of Greenwich, 2018)

D.1.2 Poster presentations

- ScotChem conference (The University of Edinburgh 2016 & University of Glasgow 2017 & The University of St Andrews 2018 & Heriot-Watt University 2019)
- CCPBioSim Conference (Derby University, July 2016 & University of Southampton, 2017)
- Joseph Black conference (The University of Edinburgh, 2017)
- Allostery RSC conference (The Royal Society, London, 2017)
- Cecam: Computational allostery (Lausanne Switzerland, 2017)
- Astbury Conversation(University of Leeds, 2018)

- Annual UCB PhD day (Royal College of Physicians London, October 2016, October 2017, October 2018)

D.1.3 Poster prizes

- 1st place poster prize:
  Joseph Black Conference
  (University of Edinburgh: 1st June 2017)

- 1st place poster prize:
  11th ScotCHEM Computational Chemistry Symposium
  (University of Glasgow: 16th June 2017)

- 1st place poster prize:
  UCB PhD day
  (London: 23rd October 2017)

- 3rd place poster prize:
  UCB PhD day
  (London: 24th September 2018)

Bibliography

[1] J. Monod and F. Jacob. "General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation". *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 26. 1961, 389.

[2] J. Monod, J.-P. Changeux, and F. Jacob. *Journal of Molecular Biology* **6**, 306 (1963).

[3] J. Monod, J. Wyman, and J.-P. Changeux. *Journal of Molecular Biology* **12**, 88 (1965).

[4] D. E. Koshland, G. Némethy, and D Filmer. *Biochemistry* **5**, 365 (1966).

[5] C. Bohr, K. Hasselbalch, and A. Krogh. *Skandinavisches Archiv für Physiologie* **16**, 401 (1904).

[6] R. Benesch and R. E. Benesch. *Biochemical and biophysical research communications* **26**, 162 (1967).

[7] K. Gunasekaran, B. Ma, and R. Nussinov. *Proteins: Structure, Function and Genetics* **57**, 433 (2004).

[8] R. Nussinov, C.-J. Tsai, and B. Ma. *Annual Review of Biophysics* **42**, 169 (2013).

[9] R. Brudler *et al. Journal of Molecular Biology* **363**, 148 (2006).

[10] Q. Cui and M Karplus. *Protein Science* **17**, 1295 (2008).

[11] A. del Sol *et al. Structure* **17**, 1042 (2009).

[12] M. Sahún-Roncero *et al. Angewandte Chemie (International ed. in English)* **52**, 4582 (2013).

[13] V. J. Hilser, J. O. Wrabl, and H. N. Motlagh. *Annual review of biophysics* **41**, 585 (2012).

[14] B. D. J. C. Kendrew *et al.* **1**, 662 (1958).

[15] T. Lengauer and M. Rarey. *Current Opinion in Structural Biology* **6**, 402 (1996).

[16] P. J. Gane and P. M. Dean. *Current Opinion in Structural Biology* **10**, 401 (2000).

[17] K. Lindorff-Larsen *et al. Science* **334**, 517 (2011).

[18] A. K. Grover. *Medical Principles and Practice* **22**, 418 (2013).

[19] R. D. Smith, J. Lu, and H. A. Carlson. *PLOS Computational Biology* **13**,ed. by E. Papaleo, e1005813 (2017).

[20] H. N. Motlagh *et al. Nature* **508**, 331 (2014).

[21] A Cooper and D. T. F. Dryden. *European Biophysics Journal* **11**, 103 (1984).

[22] R. Nussinov, B. Ma, and C.-J. Tsai. *Biophysical Chemistry* **186**, 22 (2014).

[23] R. Grünberg, J. Leckner, and M. Nilges. *Structure* **12**, 2125 (2004).

[24] P. Csermely, R. Palotai, and R. Nussinov. *Trends in Biochemical Sciences* **35**, 539 (2010).

[25] T. Wlodarski and B. Zagrovic. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 19346 (2009).

[26] I. Daidone and A. Amadei. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**, 762 (2012).

[27] C. C. David and D. J. Jacobs. *Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins*. Vol. 1084. 2014, 193.

[28] F. Sittel, A. Jain, and G. Stock. *Journal of Chemical Physics* **141** (2014).

[29] A. Altis *et al. The Journal of Chemical Physics* **126**, 244111 (2007).

[30] E. C. Dykeman and O. F. Sankey. *Journal of Physics: Condensed Matter* **22**, 423202 (2010).

[31] I. Bahar *et al. Annual Review of Biophysics* **39**, 23 (2010).

[32] A. Atilgan *et al. Biophysical Journal* **80**, 505 (2001).

[33] *Web of Knowledge*. `http://www.webofknowledge.com/`. Accessed: 20/09/2019.

[34] J. A. McCammon, B. R. Gelin, and M. Karplus. *Nature* **267**, 585 (1977).

[35] B. J. Alder and T. E. Wainwright. *The Journal of Chemical Physics* **31**, 459 (1959).

[36] J. A. Hardy and J. A. Wells. *Current Opinion in Structural Biology* **14**, 706 (2004).

[37] R. Nussinov and C.-J. Tsai. *Cell* **153**, 293 (2013).

[38] J. A. Hegler, P. Weinkam, and P. G. Wolynes. *HFSP journal* **2**, 307 (2008).

[39] J. O. Wrabl *et al. Biophysical Chemistry* **159**, 129 (2011).

[40] G. Kar *et al. Current Opinion in Pharmacology* **10**, 715 (2010).

[41] A. P. Kornev and S. S. Taylor. *Trends in Biochemical Sciences* **40**, 628 (2015).

[42] L. Di Paola and A. Giuliani. *Current Opinion in Structural Biology* **31**, 43 (2015).

[43] R. Nussinov and C.-j. Tsai. *Current Opinion in Structural Biology* **30**, 17 (2015).

[44] M. S. Marlow *et al. Nature chemical biology* **6**, 352 (2010).

[45] J. A. Maier *et al. J. Chem. Theory Comput.* **11**, 3696 (2015).

[46] A. R. Leach. *Molecular Modelling: Principles and Applications*. 2001.

[47] U. Essmann *et al. The Journal of Chemical Physics* **103**, 8577 (1995).

[48] M. Deserno and C. Holm. *Journal of Chemical Physics* **109**, 7678 (1998).

[49] I. G. Tironi *et al. The Journal of Chemical Physics* **102**, 5451 (1995).

[50] I. Fukuda and H. Nakamura. *Biophysical Reviews* **4**, 161 (2012).

[51] W. C. Swope *et al. The Journal of Chemical Physics* **76**, 637 (1982).

[52] H. J. C. Berendsen *et al. The Journal of Chemical Physics* **81**, 3684 (1984).

[53] H. C. Andersen. *The Journal of Chemical Physics* **72**, 2384 (1980).

[54] W. G. Hoover. *Physical Review A* **31**, 1695 (1985).

[55] S. Nosé. *The Journal of Chemical Physics* **81**, 511 (1984).

[56] J. Woods, C. J., Michel. *Sire/OpenMM*. 2014.

[57] M. Parrinello and A. Rahman. *Journal of Applied Physics* **52**, 7182 (1981).

[58] S Kullback and R. A. Leibler. *The Annals of Mathematical Statistics* **22**, 79 (1951).

[59] Y. Mu, P. H. Nguyen, and G. Stock. *Proteins: Structure, Function, and Bioinformatics* **58**, 45 (2004).

[60] J.-H. Prinz *et al. The Journal of Chemical Physics* **134**, 174105 (2011).

[61] C. Wehmeyer *et al. Living Journal of Computational Molecular Science* **1**, 1 (2019).

[62] H. Lu and K. Schulten. *Proteins: Structure, Function, and Genetics* **35**, 453 (1999).

[63] T. Pawson and J. D. Scott. *Trends in Biochemical Sciences* **30**, 286 (2005).

[64] P. Lahiry *et al. Nature Reviews Genetics* **11**, 60 (2010).

[65] R. M. Biondi *et al. The EMBO Journal* **20**, 4380 (2001).

[66] M. M. Keshwani *et al. Journal of Biological Chemistry* **286**, 23552 (2011).

[67] G. Manning. *Science* **298**, 1912 (2002).

[68] F. Ardito *et al. International Journal of Molecular Medicine* **40**, 271 (2017).

[69] K. C. Duong-Ly and J. R. Peterson. *Curr Protoc Pharmacolo.* **60**, 2.9.1 (2013).

[70] N. G. Anderson *et al. International Journal of Cancer* **94**, 774 (2001).

[71] A. Vultur *et al. Molecular Cancer Therapeutics* **7**, 1185 (2008).

[72] M. Deininger. *Blood* **105**, 2640 (2005).

[73] Y.-L. Lin *et al. Proceedings of the National Academy of Sciences* **110**, 1664 (2013).

[74] A. G. Gilmartin *et al. Clinical Cancer Research* **17**, 989 (2011).

[75] A. Converso *et al. Bioorganic & Medicinal Chemistry Letters* **19**, 1240 (2009).

[76] E. Park *et al. Nature Publishing Group* **22**, 703 (2015).

[77] J. Zhang *et al. Nature* **463**, 501 (2010).

[78] S. Betzi *et al. ACS Chemical Biology* **6**, 492 (2011).

[79] C. Bessa *et al. Cell Death and Disease* (2018).

[80] A. Thorarensen *et al. ACS Chemical Biology* **9**, 1552 (2014).

[81] R. Nussinov and C.-J. Tsai. *Annual Review of Pharmacology and Toxicology* **55**, 249 (2015).

[82] R. R. Yocum *et al. Proceedings of the National Academy of Sciences of the United States of America* **76**, 2730 (1979).

[83] D. C. Liebler and F. P. Guengerich. *Nature Reviews Drug Discovery* **4**, 410 (2005).

[84] R. B. Corcoran *et al. Science Signaling* **3**, ra84 (2010).

[85] W. Pao *et al. PLoS Medicine* **2**,ed. by E. T. Liu, e73 (2005).

[86] A. C. Anderson. *ACS Chemical Biology* **7**, 278 (2012).

[87] R. M. Biondi *et al. The EMBO journal* **21**, 4219 (2002).

[88] J. O. Schulze *et al. Cell Chemical Biology* **23**, 1193 (2016).

[89] B. P. Ziemba *et al. Biochemistry* **52**, 4820 (2013).

[90] R. M. Biondi *et al. EMBO Journal* **19**, 979 (2000).

[91] J. D. Sadowsky *et al. Proceedings of the National Academy of Sciences* **108**, 6056 (2011).

[92] M. M. Harding. *Acta Crystallographica Section D Biological Crystallography* **57**, 401 (2001).

[93] D. M. Jacobsen *et al. Journal of the American Chemical Society* **134**, 15357 (2012).

[94] S. J. Kerns *et al. Nature structural & molecular biology* **22**, 124 (2015).

[95] D. K. Treiber and N. P. Shah. *Chemistry and Biology* **20**, 745 (2013).

[96] J. M. Wang *et al. J. Comput. Chem.* **25**, 1157 (2004).

[97] W. D. Cornell *et al. Journal of the American Chemical Society* **117**, 5179 (1995).

[98] D. Y. D.A. Case, J.T. Berryman, R.M. Betz, D.S. Cerutti, T.E. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz and P. Kollman. *The Amber14 Software package*. University of California, San Francisco., 2015.

[99] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.8". 2015.

[100] A Sali. *Molecular medicine today* **1**, 270 (1995).

[101] *Schrödinger Release 2015-4: Maestro, Schrödinger, LLC, New York, NY, 2015.*

[102] L. G. Trabuco *et al. Nucleic Acids Research* **40**, 423 (2012).

[103] K. Busschots *et al. Chemistry & biology* **19**, 1152 (2012).

[104] L. A. Lopez-Garcia *et al. Chemistry and Biology* **18**, 1463 (2011).

[105] K. L. Meagher, L. T. Redman, and H. A. Carlson. *Journal of Computational Chemistry* **24**, 1016 (2003).

[106] M. B. Peters *et al. Journal of Chemical Theory and Computation* **6**, 2935 (2010).

[107] O. Allnér, L. Nilsson, and A. Villa. *Journal of Chemical Theory and Computation* **8**, 1493 (2012).

[108] P. Li *et al. Journal of Chemical Theory and Computation* **9**, 2733 (2013).

[109] J. Aqvist and A. Warshel. *J Am Chem Soc* **112**, 2860 (1990).

[110] L. Kamerlin. *Personal communication*. 2016.

[111] F. Duarte *et al. The journal of physical chemistry. B* **118**, 4351 (2014).

[112] T Steinbrecher, J Latzer, and D. A. Case. *J. Chem. Theory Comput.* **8**, 4405 (2012).

[113] J. Barker and R. Watts. *Molecular Physics* **26**, 789 (1973).

[114] R. T. McGibbon *et al. Biophysical Journal* **109**, 1528 (2015).

[115] S. Bhattacharya and N. Vaidehi. *Biophysical Journal* **107**, 422 (2014).

[116] M. K. Scherer *et al. Journal of Chemical Theory and Computation* **11**, 5525 (2015).

[117] F. Pedregosa *et al. SIAM Journal on Matrix Analysis and Applications* **31**, 1100 (2012).

[118] G. Varoquaux. *Non-parametric computation of entropy and mutual-information*. 2016.

[119] B. Nadler *et al. Applied and Computational Harmonic Analysis* **21**, 113 (2006).

[120] R. R. Coifman and S. Lafon. *Applied and Computational Harmonic Analysis* **21**, 5 (2006).

[121] *KL allostery GitHub scripts*. `https://github.com/michellab/KL_allostery`. Accessed: 2019-06-25.

[122] X. Gao and T. K. Harris. *The Journal of Biological Chemistry* **281**, 21670 (2006).

[123] A. Matte, L. W. Tari, and L. T. Delbaere. *Structure* **6**, 413 (1998).

[124] Z. Wang and P. A. Cole. *Methods in Enzymology* **548**, 1 (2014).

[125] M. Engel *et al. The EMBO Journal* **25**, 5469 (2006).

[126] N. K. Tonks. *FEBS Letters* **546**, 140 (2003).

[127] N. K. Tonks. *Nature Reviews Molecular Cell Biology* **7**, 833 (2006).

[128] J. S. Lazo, K. E. McQueeney, and E. R. Sharlow. *SLAS DISCOVERY: Advancing Life Sciences R&D* **22**, 1071 (2017).

[129] C. E. Gee and I. M. Mansuy. *CMLS Cellular and Molecular Life Sciences* **62**, 1120 (2005).

[130] S. Zhang and Z. Y. Zhang. *Drug Discovery Today* **12**, 373 (2007).

[131] F. Sacco *et al. FEBS Letters* **586**, 2732 (2012).

[132] Y. Shi. *Cell* **139**, 468 (2009).

[133] D. M. Virshup and S. Shenolikar. *Molecular Cell* **33**, 537 (2009).

[134] R.-j. He *et al. Acta Pharmacologica Sinica* **35**, 1227 (2014).

[135] M. J. Chen, J. E. Dixon, and G. Manning. *Science Signaling* **10**, 1 (2017).

[136] C. Fan *et al. British Journal of Cancer* **113**, 1735 (2015).

[137] N. Chalhoub and S. J. Baker. *Annual Review of Pathology: Mechanisms of Disease* **4**, 127 (2009).

[138] H. Park *et al. Journal of Biomolecular Screening* **19**, 1383 (2014).

[139] M. Porcu *et al. Blood* **119**, 4476 (2012).

[140] Y. Du and J. R. Grandis. *Chinese Journal of Cancer* **34**, 61 (2015).

[141] S. Koren and I. G. Fantus. *Best Practice & Research Clinical Endocrinology & Metabolism* **21**, 621 (2007).

[142] R. C. Tsou and K. K. Bence. *Journal of Obesity* **2012** (2012).

[143] S. M. Stanford *et al. Nature Chemical Biology* **13**, 624 (2017).

[144] J. Xu *et al. PLoS Biology* **12**, e1001923 (2014).

[145] V. Drouet and S. Lesage. *BioMed Research International* **2014**, 1 (2014).

[146] J. S. Lazo and E. R. Sharlow. *Annual Review of Pharmacology and Toxicology* **56**, 23 (2016).

[147] Z.-Y. Zhang. *Accounts of Chemical Research* **50**, 122 (2017).

[148] B. Marsh-Armstrong *et al. ACS Omega* **3**, 15763 (2018).

[149] S. M. Stanford and N. Bottini. *Trends in Pharmacological Sciences* **38**, 524 (2017).

[150] D. J. Herre *et al. PLOS ONE* **10**,ed. by K. Schröder, e0126866 (2015).

[151] P.-J. Chen *et al. Toxicology* **337**, 10 (2015).

[152] H. Liu *et al. Cancer Letters* **359**, 218 (2015).

[153] L. Lessard, M. Stuible, and M. L. Tremblay. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1804**, 613 (2010).

[154] W. Wang *et al. Biochemical and Biophysical Research Communications* **503**, 903 (2018).

[155] L. R. Bollu *et al. Clinical Cancer Research* **23**, 2136 (2017).

[156] A. M. Smith *et al. npj Regenerative Medicine* **2**, 1 (2017).

[157] C. Wiesmann *et al. Nature Structural & Molecular Biology* **11**, 730 (2004).

[158] D. A. Keedy *et al. eLife* **7** (2018).

[159] H. G. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese *et al. The Amber16 Software package*. University of California, San Francisco, 2016.

[160] M. J. Abraham *et al. SoftwareX* **1-2**, 19 (2015).

[161] M. Bonomi *et al. Nature Methods* **16**, 670 (2019).

[162] F. Noé *et al. The Journal of Chemical Physics* **139**, 184114 (2013).

[163] P. Deuflhard and M. Weber. *Linear Algebra and its Applications* **398**, 161 (2005).

[164] S. Röblitz and M. Weber. *Advances in Data Analysis and Classification* **7**, 147 (2013).

[165] S. C. L. Kamerlin, R. Rucker, and S. Boresch. *Biochemical and Biophysical Research Communications* **345**, 1161 (2006).

[166] S. C. L. Kamerlin, R. Rucker, and S. Boresch. *Biochemical and Biophysical Research Communications* **356**, 1011 (2007).

[167] G. X. Liu *et al. Acta Pharmacologica Sinica* **27**, 100 (2006).

[168] J. M. Lipchock *et al. Biochemistry* **56**, 96 (2017).